Slovenská akadémia vied
Jazykovedný ústav Ľudovíta Štúra

NLP, Corpus Linguistics, Corpus Based Grammar Research

Fifth International Conference
Smolenice, Slovakia, 25–27 November 2009
Proceedings

Editors
Jana Levická
Radovan Garabík

Tribun

2009

Slovak National Corpus
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia 2009
http://korpus.juls.savba.sk/∼slovko/

Table of Contents

Slovko (2001–2009)
Five Editions of the International Conference

Slovakia cannot boast of many linguistic events that have been organised on regular basis and focusing on one specific field. This kind of symposiums is actually rather unique, in the present-day Slovak context only three of them are known nationally: onomastic conferences are held in different parts of Slovakia, Banská Bystrica hosts conferences on communication and annual Young Linguists' Symposium covers all linguistic disciplines as well as interdisciplinary areas (in 2010 its 20th edition will be held). Moreover, this seminar once saw the early presentations of some of the pioneers of Slovak computational and corpus linguistics (Emil Páleš) and also hosted the first participants of Slovko 2001 (Karol Furdík, Jozef Ivanecký). Although there has been only 5 events named Slovko, all of them of interdisciplinary nature dealing with areas of computational and corpus linguistics, the conference gained the international character and its tradition seems to be well rooted.

2001 was the year of organization of the first Slovko conference, which was held on October 26–27 (at that time still called *Computer Processing of Slovak and Czech* and attended exclusively by Czech and Slovak lecturers and audience). This symposium represented in the first place an event organised "with the aim to improve mutual awareness and knowledge of people in Slovakia involved in the issues of computers related to the language and vice versa: language related to computers" (A. Jarošová: Malá inventúra pred hľadaním spoločného jazyka. In: *Slovenčina a čeština v počítačovom spracovaní*. Bratislava: Veda 2001, p. 7)[1]. Quoting the author and the main organiser on behalf of the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences (the second organiser being Vladmír Benko on behalf of the Faculty of Pedagogy, Comenius University), contemporary Slovakia featured only fairly isolated islets of respective activities in different scientific disciplines and theoretically-applied contexts, out of which also efforts in the area of artificial intelligence and cognitive linguistics revealed to be relevant. However, neither results of the automatised processing of Slovak language data nor relevant language data in terms of electronic corpus of texts in Slovak language were available at that time. In comparison with the Czech Republic it was a diametrically different situation since computational linguistics there had been intentionally developed as an autonomous field for more than 30 years and the 100 million token National Corpus of the Czech language has been made available on the internet in 2000. Czech lecturers could therefore "offer the Slovak professional public as well as students of linguistic and non-linguistic disciplines rather comprehensive overview of the work results in the area of Czech computer processing; this language belongs in this respect among the European and in terms of different parameters also among world's elite" (ibid). The seminar met also another of intended aims: "carry on with the scientific, pedagogic and organisational work of Ján Horecký who had been doing his best since the beginning of the 1960s to apply the

[1] Available from `http://korpus.juls.savba.sk/structure/dicts.sk.html`

principles and methods of mathematical linguistics on the material of the Slovak language" (ibid) and "who witnessed the revival of the computational linguistics in the Institute of Linguistics when he initiated in 1988–1989 a project of data base of the Slovak language, within which an idea arose to start to build a corpus" (ibid, p. 9). 14 papers altogether were presented, 5 of which by Czech authors, and consequently published in the proceedings. These proceedings were – at the same time – partial result of the Ľ. Štúr Institute of Linguistics and Faculty of Pedagogy involvement in the multinational project Trans-European Language Resources Infrastructure II – TELRI II, carried out as a coordinated action within the European Commission programme INCO-COPERNICUS in 1999–2001.

The question whether the need of mutual informing of people involved in the field of computational linguistics in Slovakia and the exchange of experience with foreign specialists was only a one-time, temporal issue or whether it had a broader context, was answered in 2003 when Vladimír Benko (as a chief organiser on behalf of the Ľ. Štúr Institute of Linguistics and Faculty of Pedagogy) held an international scientific seminar on 24–25 October, this time already called *Slovko,* subtitled as *Slavic Languages in Computer Processing*. The event was attended by 54 people interested in this area, 20 out of which were from Slovakia and the rest from abroad. This edition of Slovko saw the first presentation of the Slovak National Corpus as an actually existing electronic database of written texts covering the period 1955–2003 (in August 2003 it was made available for the public with its 26 million of tokens, in December 2003 the second version of the Corpus was released, containing 166 million of tokens) as well as the team of researchers of new department of the Ľ. Štúr Institute of Linguistics, created in mid-2002. Fledgeling corpus linguists lead by Mária Šimková presented first of all the whole of Corpus Project and then partial research solutions in segmentation, lemmatisation and morphological annotation of texts of the Slovak language. The first Slovak morphological tagset presented at Slovko was shortly afterwards publicly examined and reviewed by Slovak and foreign specialists and at the beginning of 2004 started a manual annotation of selected texts of the Slovak National Corpus.

Due to the fact that no Slovak university or college offered either an autonomous course of study or a specific seminar of computational and corpus linguistics, the Slovak National Corpus department as a chief investigator of the *State Research Programme: Integrated Computational Processing of the Slovak Language for Linguistic Research Purposes* organised regular lectures and seminars focused on these two disciplines. A part of it was published in the proceedings entitled Insight into the Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006. 208 p.[2] As a natural conclusion of these and also of some other activities was the organisation of Slovko on 10–12 November in 2005 and receiving the Slovak Academy of Sciences Prize for Building Infrastructure in Science on 11 November 2005. The third international seminar Slovko had already a programme and organizing committee and thanks to aroused interest it was also extended to computer treatment of Slavic and East European languages. It was attended by more than 60 interested researchers, which has represented the highest number of participants of one edition of Slovko. Altogether 29 papers were presented in more-or-less homo-

[2] Available from `http://korpus.juls.savba.sk/publications/`

genous sessions: spoken corpora, speech analysis and synthesis, computer lexicography, parallel and historical corpora, terminology, e-learning, which were later on published in the proceedings *Computer Treatment of Slavic and East European Languages* (Ed. R. Garabík. Bratislava: Veda 2005. 246 p.).

Slovko in 2007 (explicitly held on October 25–27) was again organised by the Slovak National Corpus team (namely Radovan Garabík, Jana Levická, and Mária Šimková), this time as a solid fourth edition of a biannual international conference focused on NLP and computational lexicography and terminology. More than 50 participants had an opportunity to attend 37 lectures and presentations; the proceedings entitled *Computer Treatment of Slavic and East European Languages 2007* (Ed. J. Levická, R. Garabík. Bratislava: Tribun 2007. 318 p.) were available upon arrival at the conference. Beside the topics that have been fundamental for every Slovko (corpus development including spoken corpora: data collection, annotation and processing), the foreground was reserved for theoretical issues of computational lexicography and terminography, bilingual lexicography and terminography, co-occurrence analysis and pertinent collocations of lexicographical and terminographical relevance, new methods in data extraction and terminology mining from corpora, terminology databases and terminology management systems and so forth. Contributions of Slovak authors covering basic as well as applied research pointed to revolutionary changes that Slovak computational and corpus linguistics in particular had seen in the recent period. For the sake of the example, we can mention several results of the Slovak National Corpus team work: on January 1st 2007 began the second phase of the realisation of the State Research and Development Programme: *Building the Slovak National Corpus and Integrated Computational Processing of the Slovak Language for the Linguistic Research Purposes*, the 350 million token fully lemmatised and morphologically annotated general corpus was made available for public. Beside other products (database of digitised lexicographic and other linguistic works, parallel corpora), highest priority has been assigned to a partial project of the Slovak National Corpus, the Slovak terminology database, which was presented by its chief investigator Jana Levická, as well as the Slovak Spoken Corpus whose technical features were shown by (in cooperation with Milan Rusko) R. Garabík.

We decided to hold Slovko 2009 in the Congress Centre Smolenice of the Slovak Academy of Sciences to celebrate sort of a jubilee edition since it is the fifth in a row (all the previous editions took place in Bratislava but in different venues). Originally, the conference was to have been focused on parallel corpora, however, since the Institute of the Czech National Corpus, Charles University, Prague, organised a conference with the same topic on 17–19 September 2009 within their InterCorp Project, we sought to come up with a different key issue that could be of immediate interest in the Central-European context. Since the Institute of the Czech Language, Academy of Sciences of the Czech Republic ceased to organise regular conferences entitled *Grammar & Corpora* (hosted by František Štícha), the main conference topic was in fact right at hand: corpus linguistics, more precisely corpus based linguistic research. In this respect this year's Slovko has once again gathered more Slovak and Czech participants, which can be viewed also as a reminder of the beginnings of Slovko. Presented contributions point to other topics and, at the same time,

give evidence of the progress in this field: majority of them do not deal with a development of standard corpora of written or spoken texts but rather with the tools of text analysis, building, and usage of language resources, linguistic components of information systems, computer-aided translation systems, localisation and lexicography, computer-aided language learning etc. Similarly as the organisers of the first Slovko participated in international projects, the Slovak National Corpus team and Slovak computational and corpus linguistics have been involved in several international cooperation projects, especially in the 7th framework programme (EU FP7 INF 211983 Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources).

Five editions of Slovko reflect the corpus and computational linguistics in Slovakia, Czech Republic, as well as in neighbouring and more distant countries, namely Austria, Belgium, Bulgaria, France, Croatia, Hungary, Germany, Norway, Netherlands, New Zealand, Poland, Russia, Slovenia, Spain, Ukraine have had their participants so far. Approximately 130 participants altogether from 17 countries have attended Slovko conference (however, some of them have participated repeatedly, on regular basis). Overall, Slovko conferences hosted almost 300 participants that could listen to and/or read nearly 150 papers. We would like to express our gratitude for excellent cooperation to all members of programme committees and abstract peer reviewers. We also thank to all participants for creating fruitful working and social atmosphere, to Czech colleagues in particular for they have been showing us support from the beginning and attended regularly all Slovko's editions. We appreciate very much the work of all members of each organising committee and hope that the flag will be handed on.

Mária Šimková
Jana Levická

Slovko (2001 – 2009)
Päť ročníkov medzinárodného podujatia

V rámci lingvistických podujatí na Slovensku nemáme veľa takých, ktoré by sa uskutočňovali pravidelne a boli zamerané na jednu oblasť. Takéto podujatia sú naozaj skôr výnimočné, v súčasnom slovenskom kontexte sú širšie známe asi tri: na rôznych miestach Slovenska sa konávajú onomastické konferencie, v Banskej Bystrici konferencie o komunikácii a celú lingvistiku, ale aj interdisciplinárne oblasti zahŕňa každoročné Kolokvium mladých jazykovedcov (v r. 2010 sa už chystá 20. ročník), na ktorom sa ako mladí kedysi prezentovali aj jedni z priekopníkov slovenskej počítačovej a korpusovej lingvistiky (E. Páleš) a prví účastníci Slovka 2001 (K. Furdík, J. Ivanecký). Podujatí s názvom Slovko, ktoré sú zamerané interdisciplinárne na oblasti počítačovej a korpusovej lingvistiky, bolo zatiaľ len päť, no ukazuje sa, že sa založila dobrá tradícia s medzinárodným presahom.

Keď sa v r. 2001 pripravovalo a 26. – 27. októbra aj uskutočnilo prvé Slovko (ešte pod názvom Slovenčina a čeština v počítačovom spracovaní a za účasti výlučne českých a slovenských prednášateľov i poslucháčov), išlo predovšetkým o podujatie organizované „so zámerom zlepšiť vzájomnú informovanosť ľudí zaoberajúcich sa na Slovensku problematikou počítačov vo vzťahu k jazyku a naopak, jazyka vo vzťahu k počítačom“ (A. Jarošová: Malá inventúra pred hľadaním spoločného jazyka. In: Slovenčina a čeština v počítačovom spracovaní. Bratislava: Veda 2001, s. 7)[1]. Ako konštatovala autorka a hlavná organizátorka seminára za Jazykovedný ústav Ľ. Štúra SAV (druhým organizátorom bol V. Benko za Pedagogickú fakultu Univerzity Komenského), v tejto oblasti boli v tom čase na Slovensku značne izolované ostrovčeky aktivít prebiehajúcich v rôznych vedných odboroch a teoreticko-aplikačných kontextoch, z ktorých sa ako relevantné ukazovali aj aktivity v oblasti umelej inteligencie a kognitívnej lingvistiky. Neboli však k dispozícii výsledky automatizovaného spracovania slovenských jazykových dát ani relevantné jazykové dáta v podobe elektronického korpusu textov slovenského jazyka. Išlo o diametrálne odlišnú situáciu v porovnaní s Českou republikou, kde sa počítačová lingvistika cieľavedome budovala ako samostatný odbor viac ako tridsať rokov a národný korpus českého jazyka bol v r. 2000 zverejnený na internete v rozsahu 100 miliónov tokenov. Českí prednášatelia preto mohli „poskytnúť slovenskej odbornej verejnosti a študentom z lingvistických aj nelingvistických odborov ucelenejší pohľad na výsledky práce v oblasti počítačového spracovania češtiny, ktorá patrí v tomto smere medzi európsku a vo viacerých parametroch aj medzi svetovú špičku“ (tamže). Seminár splnil aj ďalší zo svojich cieľov: „nadviazať na vedeckú, pedagogickú a organizačnú prácu Jána Horeckého, ktorý sa od začiatku 60. rokov 20. storočia usiloval uplatňovať princípy a metódy matematickej lingvistiky na materiáli slovenského jazyka“ (tamže) a „ktorý stál aj pri revitalizácii počítačovej lingvistiky v Jazykovednom ústave, keď v rokoch

[1] Dostupný aj z `http://korpus.juls.savba.sk/structure/dicts.sk.html`

1988 – 1989 pripravil projekt bázy dát slovenského jazyka, v rámci ktorej sa začalo uvažovať aj o budovaní korpusu“ (tamže, s. 9). Na seminári odznelo a v zborníku z neho bolo publikovaných 14 príspevkov, z toho 5 od českých autorov. Publikovaný zborník zároveň predstavoval jeden z výsledkov účasti JÚĽŠ SAV a PdF UK v mnohonárodnom projekte Trans-European Language Resources Infrastructure II – TELRI II, ktorý sa ako súbor coordinated action uskutočnil v rámci programu Európskej komisie INCO-COPERNICUS v r. 1999 – 2001.

Otázka, či potreba vzájomnej informovanosti pracovníkov v oblasti počítačovej lingvistiky na Slovensku a výmena skúseností s odborníkmi zo zahraničia bola iba jednorazová, dočasná, alebo má širší kontext, bola zodpovedaná v r. 2003, keď V. Benko (ako hlavný organizátor za JÚĽŠ SAV i PdF UK) uskutočnil 24. – 25. októbra medzinárodný odborný seminár už pod názvom Slovko, v podtitule špecifikovaný na Slovanské jazyky v počítačovom spracovaní. Podujatia sa zúčastnilo 54 záujemcov o túto oblasť, z toho 20 zo Slovenska a 34 zo zahraničia. Na tomto Slovku sa prvýkrát prezentoval Slovenský národný korpus – ako už reálne existujúca elektronická databáza písaných textov z r. 1955 – 2003 (od augusta 2003 verejne prístupná v rozsahu 26 miliónov tokenov, v decembri 2003 bola zverejnená druhá verzia v rozsahu 166 miliónov tokenov) i ako kolektív pracovníkov nového oddelenia JÚĽŠ SAV, založeného v podstate v polovici r. 2002. Začínajúci korpusoví lingvisti pod vedením M. Šimkovej prezentovali jednak celý korpusový projekt, jednak čiastkové riešenia pri segmentácii, lematizácii a morfologickej anotácii textov slovenského jazyka. Prvá podoba slovenského morfologického tagsetu predložená na Slovku bola zakrátko osobitne verejne oponovaná domácimi aj zahraničnými odborníkmi a začiatkom r. 2004 sa rozbehla ručná morfologická anotácia vybraných textov SNK.

Vzhľadom na to, že na žiadnej vysokej škole na Slovensku neexistoval samostatný odbor a ani sa nevyučoval samostatný predmet počítačovej a korpusovej lingvistiky, oddelenie SNK JÚĽŠ SAV ako riešiteľ štátnej úlohy Integrated Computational Processing of the Slovak Language for Linguistic Research Purposes organizovalo pravidelné prednášky a semináre z tejto oblasti. Časť z nich bola publikovaná v zborníku Insight into the Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006. 208 s.[2] Prirodzeným vyústením týchto a ďalších aktivít bolo zorganizovanie Slovka 10. – 12. novembra 2005 a prevzatie Ceny SAV za budovanie infraštruktúry pre vedu 11. novembra 2005. Tretí medzinárodný seminár Slovko už bol obohatený o vedecký výbor a organizačný výbor a vzhľadom na prejavený záujem bol rozšírený na Slovanské a východoeurópske jazyky v počítačovom spracovaní. Zúčastnilo sa ho vyše 60 záujemcov, čo bol a stále je najvyšší počet účastníkov jedného Slovka. Celkovo odznelo 29 príspevkov vo viac-menej homogénnych blokoch: hovorené korpusy, analýza a syntéza reči, počítačová lexikografia, paralelné korpusy, historické korpusy, terminológia, e-learning a pod. Tieto vyšli v zborníku Computer Treatment of Slavic and East European Languages (Ed. R. Garabík. Bratislava: Veda 2005. 246 s.).

[2] Dostupný aj z `http://korpus.juls.savba.sk/publications/`

Slovko v r. 2007 (25. – 27. októbra) organizoval kolektív SNK (najmä R. Garabík, J. Levická, M. Šimková) už ako stabilizovanú 4. bienálnu medzinárodnú konferenciu so zameraním na NLP a počítačovú lexikografiu a terminológiu. Vyše 50 účastníkov si malo možnosť vypočuť 37 príspevkov a prezentácií, zborník Computer Treatment of Slavic and East European Languages (Ed. J. Levická, R. Garabík. Bratislava: Tribun 2007. 318 s.) mali všetci k dispozícii priamo na podujatí. Okrem tém, ktoré sú ako základné súčasťou každého Slovka (tvorba korpusov vrátane hovorených: zber dát, anotácia a spracovanie), sa do popredia dostali teoretické otázky komputačnej lexikografie a terminografie, bilingválna lexikografia a terminografia, kookurenčná analýza a lexikograficky alebo terminograficky relevantné kolokácie, nové metódy v extrahovaní dát a získavanie terminológie z korpusov, terminologické databázy a systémy terminologického manažmentu a pod. Príspevky slovenských autorov, pokrývajúce základný aj aplikovaný výskum, poukázali na prevratné zmeny, ktoré slovenská počítačová a najmä korpusová lingvistika zaznamenala za posledné obdobie. Napr. z produkcie SNK, ktoré 1. januára 2007 začalo 2. etapu riešenia štátnej úlohy *Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku*, bol verejnosti k dispozícii hlavný korpus v rozsahu 350 miliónov tokenov s plnou lematizáciou a morfologickou anotáciou na báze vlastného tagsetu. Okrem ďalších produktov (databáza elektronických lexikografických a iných lingvistických diel, paralelné korpusy) sa prioritným čiastkovým projektom SNK stala Slovenská terminologická databáza, ktorú na konferencii predstavila jej hlavná riešiteľka J. Levická, a budovať sa začal aj Slovenský hovorený korpus, ktorého technické parametre prezentoval (v spolupráci s M. Ruskom) R. Garabík.

Slovko 2009 sme ako polojubilejné, piate v poradí umiestnili do Kongresového centra SAV v Smoleniciach (všetky predchádzajúce boli v Bratislave, a to na viacerých miestach). Hlavným zameraním mali byť pôvodne paralelné korpusy, ale keďže Ústav Českého národného korpusu FF UK v Prahe organizoval takto profilovanú konferenciu v súlade s cieľmi svojho projektu InterCorp 17. – 19. septembra 2009, hľadali sme inú hlavnú tému, ktorá by mohla byť aktuálnou v stredoeurópskom priestore. A keďže Ústav pro jazyk český AV ČR prestal organizovať pravidelné konferencie Grammar & Corpora (pod vedením F. Štíchu), priam sa nám ponúkla ako hlavná téma korpusová lingvistika, resp. korpusovo zamerané lingvistické výskumy. V tejto súvislosti je na tohtoročnom Slovku opäť viac slovenských a českých účastníkov, čo je aj istá pripomienka východiskovej situácie Slovka 2001. Posun v tejto oblasti nielen v našich krajinách naznačujú aj ďalšie prezentované témy: už to nie sú otázky budovania štandardného korpusu písaných či hovorených textov, ale viac nástroje textovej analýzy, tvorba a využitie jazykových zdrojov, lingvistické zložky informačných systémov, preklad s počítačovou podporou, lokalizácia a lexikografia, didaktika vyučovania cudzích jazykov s počítačovou podporou a pod. A tak ako organizátori prvého Slovka boli členmi medzinárodných projektov, aj Slovenský národný korpus a slovenská počítačová a korpusová lingvistika sa v súčasnosti prezentuje vo viacerých medzinárodných spoluprácach, predovšetkým v 7. RP (EU FP7 INF 211983 s nazvom Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources).

Päť ročníkov Slovka je odrazom vývinu počítačovej a korpusovej lingvistiky na Slovensku, v Čechách, v okolitých i vzdialenejších krajinách, z ktorých sú doteraz zastúpené Belgicko, Bulharsko, Francúzsko, Gruzínsko, Holandsko, Chorvátsko, Maďarsko, Nemecko, Nórsko, Nový Zéland, Poľsko, Rakúsko, Rusko, Slovinsko, Španielsko, Ukrajina. Zahraničných účastníkov bolo na všetkých konferenciách Slovko spolu asi 130 (nie sú to unikátne výskyty, viacerí sú našimi pravidelnými hosťami) zo 17 krajín. Spolu so slovenskými účastníkmi sme privítali na týchto podujatiach takmer 300 záujemcov, ktorí si mali možnosť vypočuť a/alebo prečítať takmer 150 príspevkov. Za vynikajúcu spoluprácu ďakujeme všetkým členom vedeckého výboru a posudzovateľom abstraktov. Ďakujeme všetkým účastníkom za vytváranie tvorivej pracovnej i spoločenskej atmosféry, osobitne českým kolegom, ktorí nás od začiatku podporovali a pravidelne sa zúčastňovali aj Slovka. Ďakujeme všetkým členom doterajších organizačných výborov a dúfame, že štafeta pôjde ďalej.

Mária Šimková
Jana Levická

Scientific Text Corpora as a Lexicographic Source

Larisa Belyaeva

Herzen State Pedagogical University of Russia
Saint-Petersburg, Russia

Abstract. Nowadays applied lexicography is one of the special branches of applied linguistics, its task is creation and updating different automated and automatic dictionaries and databases, which are problem and subject oriented. Completeness and adequacy of lexicographic systems to a considerable extent determine the level and reliability of information and knowledge extraction from the texts of various composition, structure and assignment.
Modern approach to dictionary creation assumes preliminary formation and use of parallel corpora of modern texts, which can be considered as a database for solving not only research tasks, but practical lexicographic tasks as well. If we use a full-text parallel corpora as a lexicographic base it is necessary to expand them with a corpora of machine translation results, analysis and comparison of these corpora will make it possible to allocate such lexical units, which should be considered as dictionary entries. The main problem is to establish the boundaries and structures of these lexical units.

1 Applied lexicography and scientific texts

Nowadays applied lexicography is one of the special branches of applied linguistics, its task is creation and updating different automated and automatic dictionaries and databases, which are problem and subject oriented. Completeness and adequacy of lexicographic systems to a considerable extent determine the level and reliability of information and knowledge extraction from the texts of various composition, structure and assignment.

Modern approach to dictionary creation assumes preliminary formation and use of parallel corpora of modern texts, which can be considered as a database for solving not only research tasks, but practical lexicographic tasks as well. Written text corpora, as a rule, include the texts as they are, as well as text layouts: format boundaries and features, parsing results necessary for establishing morphological characteristics of lexical units. These texts can be used serve for concordance creation, word and collocation lists in case of monolingual corpora, as well as for creation multilingual lexicons and concordance if we have parallel corpora.

If we use full-text parallel corpora as a lexicographic base it is necessary to expand them with a corpora of machine translation results assumes, analysis and comparison of these corpora will make it possible to allocate such lexical units, which should be considered as dictionary entries. The main problem is to establish the boundaries and structures of these lexical units.

Noun phrases are objects of special research in both theoretical and applied aspects. Such phrases are functionally equivalent to a word, but at the same time they represent a convolution of a sentence, i.e. they are rather units of syntax, not lexicon. Thus we can assume that internal structure of a noun phrase correlate with internal dependencies structure of the appropriate sentence. The problem is to find a procedure or approach to recognize this structure in a convolution.

One of the most serious problems of English scientific text analysis and machine or human translation is determination of dependency structure in noun phrases (NP). The problem is related to the fact than when translating from English to any inflectional language we should know the relation structure between the NP components.

In scientific text simple noun phrases are multicomponent units with large number of attributive elements in preposition to the NP head. Being dependent members of a sentence, these phrases form one syntactic group with its head, syntactic function of which coincides with syntactic function of the phrase as a whole. The report investigates the possibility of using corpora information for solving the problem of NP structure recognition and translation.

Written text corpora, as a rule, include the texts as they are, as well as text layouts: format boundaries and features, parsing results necessary for establishing morphological characteristics of lexical units. These texts can be used for concordance creation, word and collocation lists in case of monolingual corpora, as well as for creation multilingual lexicons and concordance if we have parallel corpora.

2 Noun groups in a scientific text

During the translation process the text analysis is based on formal parsing and semantic analysis. Both of these processes are based on our possibility to understand the surface structure of a sentence and semantic relations between its components.

Noun phrases (NP) are the objects of special research in both theoretical and applied aspects. Such phrases are functionally equivalent to a word, but at the same time they represent a convolution of a sentence, i.e. they are rather units of syntax, not lexicon. So we can assume that internal structure of a noun phrase correlate with internal dependencies structure of the sentence. The problem is to find a procedure to recognize this structure in a concise form of a NP.

Thus one of the most serious problems of English-Russian scientific text analysis and machine or human translation is determination of dependency structure in NPs. The problem is related to the fact than when translating from English to any inflectional language we should know the relation structure between the NP components. In scientific text simple (without preposition) NPs are multi-component units with large number of attributive elements in preposition to the NP head element. Being dependent members of a sentence, these phrases form one syntactic group with its head element, syntactic function of which coincides with syntactic function of the phrase as a whole.

In machine translation procedure the structure of each NP and its boundaries are to be determined at the sentence analysis step, thus the task of NP translation is to be performed in the framework of the following operations:

- Establishing the head element of the English noun phrase;
- Establishing semantic and syntactic structure of the English NP;
- Finding semantic, syntactic and lexical structure of the Russian NP;

Since a NP is a sentence convolution, a compression of this structure, such external simplification of both the structure and the form causes the NP semantic complication. The markers of relations between actual components and types of relations between elements, which sentence shows with the help of different means, are absent in the English NP.

Basic noun phrases in English are two-element combinations with a head noun, frequency of which in scientific text exceeds frequency of three-element combinations in three times. (see Tables 1-3).

Rang	NP type	NP frequency	Accumulated frequency	%
1	T + N	2253	2253	23,07
2	A + N	1584	3837	39,29
3	N1 + N2	1368	5205	54.32
4	A + N1 + N2	485	5690	58.26
5	H + N	241	5931	60,74
6	M/S + N	215	6146	62„93
7	E + N	208	6354	65,06
8	N1/A + N2	199	6553	67,09
9	S + N	195	6748	69,09
10	N1 + N2 + N3	173	6921	70.87

Table 1. Fragment of frequency list of English NP structures

No	NP length	Number of different models	NP frequency
1	2	1516	3457
2	3	674	1053
3	4	207	380
4	5	51	61
5	6	20	164
6	7	2	2
7	8	5	5
Total		2475	5122

Table 2. Frequency of English noun phrase length in technical texts (subject domain “Seismic protection”)

Model number	Model	Length	Frequency	Number of different NP
1	A+N	2	1474	748
2	A+PII+N	3	9	6
3	N1+N2	2	1407	530
4	N1+N2+N3	3	248	128
5	A/N1+N2	2	71	47
6	N1+G/N2	2	24	18
7	A1+A2/N1+N2	3	10	10
8	A+G/N2 +N2	3	7	4
9	A1+A2+N	3	151	104
10	A+N1+N2	3	292	172
11	PII+A+N	3	25	20
12	PII+N	2	170	73
13	A1+N1+A2/N2+N3	4	3	2
14	A1+C+A2+N	4	15	9
15	A1+A2/N1+N2+N3	4	6	7

Table 3. Models of noun phrases in technical texts

However external simplicity of the most frequent English NP structures is misleading. The fact is that this simplicity could be the result of initial noun phrase or sentence compression. Such compression, formal simplification of NP structure leads to its semantic complication.

Pursuant to these, formation of noun phrases in a real text is based on either merging noun phrases and separate lexical units in a new, more complicated nominative construction, or on condensing multi-component NPs at the expense of deletion of the units which are implicitly obvious.

Formation of a multi-component NP in a text is realized in any of two ways depending on the type of nomination: either as a process of step-by-step complication and specification of the object nomination (gradual complication of a noun phrase with addition of its head element characteristics), or as a process of sequential convolution. This process is realized successively on several levels:

Level 1: transfer from a complex noun phrase to a simple one due to element inversion.

Level 2: elimination of component duplication in a new NP.

Level 3: coordination of semes and elimination of components with duplicated semes.

Referential status of noun phrases in scientific text permits us assume that author's attitude on information transfer and its understanding requires explication of relations within the text. Analysis of texts in different subject domains had shown, that occurrence in the text a NP with length more than 2 elements is followed by occurrence a

2-compound NP in the nearest context, within the limits of 2-3 sentences or combination of title, key words and abstract. Hence, at human translation we can use this situation as a key for structure diagnostics. At MT we need to create a special text translation memory.

The peculiarities of NP formation in the text are to be analyzed as in the following example:

Connection of two two-element NPs into a new one which results in occurrence of

- four-component NP, the structure of which depends on the structure of merging NPs, for example, if two groups of adjective + noun type merger the NP which plays the role of an attribute is embedded in the position of the head element of the first NP attribute:

indirect method + seismic analysis ⇒ indirect seismic analysis method
adult learner + second language ⇒ adult second language learner

- three-component NP in case, when one of the elements in two initial NP coincides, for example
  mental processing + processing operation ⇒
  mental processing operation
  with establishing direct relations between (in this case) adjective *mental* and noun *processing*,
- three-component NP in case, when semantics of one of the NP elements is supported as a part of a new noun phrase by the semes of other NP components, for example, merger of NPs
  communicative method + language learning
  results in occurrence of a new noun phrase
  communicative language learning
- three-component NP in case, when semantics of one of the elements is implied as a part of a new noun phrase at the expense of extralinguistic information of the domain, for example, merger of NPs
  seismic stability + direct analysis
  results in formation of the NP
  seismic stability direct analysis,
  which in the text may be convoluted up to three-component NP
  seismic direct analysis

The cases of noun phrase transformation considered here under the condition of text coherence and cohesion do not show all possible variants of their development in a text, however, they give the basis for consideration of possible translation of a noun phrase with high degree of structure compression. Besides the research conducted permits to show, that exactly two-element noun phrases present special difficulties at their analysis and translation.

To solve the problem of such NP translation we can see only two approaches which can be used both in human and machine translation MT).

The first approach includes modeling the knowledge base of the domain in question (in the framework of a MT system) or appealing to such factual knowledge of the translator. In case of MT this approach is based on vast investigations of the possible relations between both the main concepts of the domain and the items of the linguistic data base. Creation of such a thesaurus or a semantic net is not only extremely laborious but space-consuming. But the most serious disadvantage of this approach is that an unambiguous solution of the problem sometimes can't be achieved. For example, for a noun phrase *constant amplitude deformation cycle* a semantic network would show relations between the nodes *constant* and *amplitude*, *constant* and *deformation*, *constant* and *cycle* and it is impossible to use this information to establish the dependencies structure of the NP both in human and machine translation.

The second approach could be more formal: we can use the information, which can be received on the basis of the whole text analysis. This approach seems more expedient as it is based on the formal indications of the author's intentions which are reflected both in the text structure and in the composition of different NP with the same constituents.

Investigations of text structure in terms of NP composition in different subject domains (medicine, seismic isolation, space systems, power plants construction, language teaching etc) had shown that dependency structure of NP with 3 or more constituents can be obtained from the nearest context: a 2-component NPs would show the accurate relations relevant for this special text.

This means that for English-Russian MT we need a special means for noun phrase analysis and transfer within the text boundaries, something like Text Translation Memory, which could store the history of NP development.

Noun phrases are objects of special research in both theoretical and applied aspects. Such phrases are functionally equivalent to a word, but at the same time they represent a convolution of a sentence, i.e. they are rather units of syntax, not lexicon. Thus we can assume that internal structure of a noun phrase correlate with internal dependencies structure of the appropriate sentence. The problem is to find a procedure or approach to recognize this structure in a convolution.

One of the most serious problems of English scientific text analysis and machine or human translation is determination of dependency structure in noun phrases (NP). The problem is related to the fact than when translating from English to any inflectional language we should know the relation structure between the NP components.

In scientific text simple noun phrases are multicomponent units with large number of attributive elements in preposition to the NP head. Being dependent members of a sentence, these phrases form one syntactic group with its head, syntactic function of which coincides with syntactic function of the phrase as a whole. The report investigates the possibility of using corpora information for solving the problem of NP structure recognition and translation.

The Corpus of Georgian Dialects

Marine Beridze and David Nadaraia

Arn. Chikobava Institute of Linguistics
Georgian Academy of Sciences, Tbilisi, Georgia

Abstract. The Georgian Language world is represented by three Kartvelian languages and more than 25 dialects. The project "The Linguistic Portrait of Georgia" is aimed to associate the problem of documentation and researching of the Georgian dialects to the achievements of the corpus linguistics. The corpus is now under development, in which quite vast textual collection is integrated. It involves all the dialectal texts published during the last 100 years, archive material obtained in all the dialectological field expeditions which took place in the second half of the last century, and additionally dialectal texts recorded by the project group in Georgia and its neighboring countries (Azerbaijan, Iran).

1 Introduction

The idea of corpus presentation and learning of the dialectal data by means of modern technologies occurred in early researches based on the principles of corpus linguistics, and seemed quite natural. Dialectology in its nature has always been more "corpusal" among the other branches of Linguistics (as based on the specially gathered textual material) and with its methods of description and investigation it has always "required" more implementation technological progress achievements – Transition from dialectographic paper to the electro-digital carriers was the result of following the technical progress step by step. At the very beginning of corpus Linguistics development, most of the methods of the new discipline (e.g. preparation of texts collection, experience in textual and meta-textual annotation and so on) were familiar for dialectology. Hence, relation between dialectal data and using the computer system became available from the 70-ies of the last century (Gordon R. Wood 1969).

When "the great corpus construction works" – creation of national languages corpora – began, this became important trigger to develop corpus dialectology. According to the aim of the creators, several conceptions of processing of the dialectal material data shaped out. Some researchers represent the dialectal data as "deviations" from the literary language; that is why they describe only that part of the data which differs from the "standard" language; others consider a dialect as a model of the whole cultural-communicational area and treat each component more carefully (Kryuckova, Goldyn 2008); quite different is the approach of those investigators and creators, who create the corpora in the format of "sustaining the languages under the risk of disappearing", aimed to documental investigation of small data of such languages, and so on.

The corpus of the Georgian dialects represents an attempt to create the most likely model of the diverse language portrait of Georgia. The modern literary Georgian is based on Kartvelian languages and their dialects; thus, it is impossible to investigate the Georgian Language without considering data of the dialects. It should be noted that in addition to the scientific importance (linguistic and culturological studies), such corpora seem one of the best means to prevent the threats to the existence of some languages and language products. The new millennium set this problem in new light. The powerful wave of globalization rapidly changes the conceptions having been established during centuries, the concept of the language security among them. The State status, media, the education system and other social components can no more provide the capable means of the language security "instituted" in the XX century as guarantors. Today, any language, not having proper reflection of its characteristics stored in computer system, or strong links with the global language space or say it otherwise – do not have a translation system, strong textual massive and tools of corpus investigation based on this, or do not have hyper textual thesauri for synchronic and diachronic analysis of the lexis fund and so on, any such language is under great risk of disappearing.

Unfortunately, in those years when the other countries actively strove to meet one of the main challenges of the epoch – to document and investigate language events documentation and investigation with modern technologies i.e. to develop corpus linguistics, our country was dwelling in the chaos of the unannounced war and computational linguistics was "developing" only symbolically, through the separate conceptions of language modeling written manually, at the side of the extinguished computers.

Consequently, our project "Linguistic portrait of Georgia" did not result from the logic and natural development of the corpus linguistics as this usually happened in another countries (electronic textual databases/libraries + achievements of the computational linguistics, namely successes reached in the processing of the natural languages: conceptions of morphological and syntactic tagging, which were realized through the invention of automatic instruments of tagging…). Our work was not research, based on theoretical knowledge, it was rather a "instinctual reaction" of the group of linguists on the challenge, assuming this challenge with all means available for them.

After the fact of entering computer technologies in our lives became actual, the idea appeared about creation large textual massive of the Georgian dialects, which would be not only the object of studies but efficient tool for those studies. Our goal was to process great textual massive using computer technologies and creation of new communication area, where the model of the language, developed both in time (XX century) and space viewpoint, and most likely to the actual situation.

We confess that we became familiar with great achievements of the corpus linguistics in the world only after we began this work.

2 The principles of compiling the corpus of the Georgian dialects.

2.1 The foundation principles of the main textual base

The Georgian language is represented by three Kartvelian languages and more than two decade of their dialects. The literary Georgian itself is represented by more than 16 dialects. Three of them are distributed outside Georgia: Fereidanian (Iran, 35–40 thousand language speakers), "Turkish Georgian" (in its turn quite diverse, including some dialectal layers), and Ingilo (Azerbaijan territory, speech of Georgians (Christians and muslims) by origin, about 15 thousand speakers).

Scientific documentation and scientific study of the Georgian dialects began from the very beginning of the XX century, though the early attempts of marking dialectal entities can be seen in the ancient manuscripts (V–VII centuries). Starting from the 20s of the last century, collecting and studying of the dialectal data became one of the priorities of the Georgian linguists scientific work. It was absolutely clear that complex grammatical system of the Georgian could not be studied without investigation of its dialects. The results of the dialectological studies were depicted in the fundamental works of the Georgian linguists; for example, "Khevsurian Poetry" collected and published by Akaki Shanidze (Khevsureti is one of the Georgian highland ethnographic region, nowadays almost emptied of its population, hundreds of which are scattered in different regions of the country). At the end of the 30-ies of the last century, great Georgian linguist and historian Ivane Javakhishvili arranged "blitz-expedition" almost in all regions of Georgia (except Iran and Turkey), to collect the dialectal data concerning home industry and handcrafts. The data was written down manually but strictly in accordance with the main requirements of the dialect peculiarities. This expedition was distinguished because it was conducted in two months and enveloped almost the whole area of the language spreading. The data of this expedition is unique by its vast content and the possibility to represent this content from the time and space viewpoint.

After the II World war was ended, for merely linguistic aims great campaign of texts collecting started, which resulted in publishing the collection "Georgian Dialectology", in 1961 (number of pages – 600). After this separate dialectal texts were published and then, in the 80-ies, a number of expeditions were conducted almost in every region of Georgia; these expeditions were equipped with several technical tools such as magnetic audio careers to fix the peculiarities of the Georgian language in the whole area of its spreading. In the beginning of the new millennium digital video-devices fixed hundreds of hours of the dialectal texts. Nowadays the new texts collected by our project group are processed and integrated in the corpus continuously, stage by stage.

We never had the "luxury" to begin our work based on the ready framework of the electronic libraries or databases, and we could not wait till these were established, and only then begin to think on the principles of text massive creation, on development of

theoretical concepts, provided that text massive creation is the indispensable pre-condition for development of the corpus linguistics and computational linguistics in general (Ershov 1983).

We are building up the structure of the corpus, the principles of tagging and texts collection all simultaneously. Furthermore, we have to carry out all kinds of work at one time instead doing it in stages, such as: collecting the material, decoding (discourse transcription), unification, digitalization, integration in the corpus (with appropriate tagging) and so on.

2.2 Dialectal lexis in the corpus (database of dictionaries and raw material for dictionaries)

As it is known from Ziph's law, a language is a great collection of rarely acknowledged occurrences (Kutuzov). This law was refined and approved on the basis of the corpus linguistics. Certainly, it must always be considered when compiling the dialectal corpus. How much we will try 'thematically motivate" a speaker in the process of collecting dialectal data, all the same it is impossible to represent complete realization of the language inventory, so the perfect model of the dialect can not be achieved. That is why we decided to integrate dialectal dictionaries and materials for the dictionaries in the corpus, with the status of the text. This will ensure full representation of the words in the corpus and on the other hand, it will give additional context for realization of a lexeme (through the textual illustrations) and besides, will assist in preparation of the cumulative dialectal dictionary. The scheme of integration of dictionaries in the corpus is as follows: a dictionary itself is a sub corpus, the components of which are integrated in the corpus structure with different "status": dictionary word-forms and the rest of the entry are related to each other as usually is a word and its context in the concordance; as for the illustration material, it is organized in the corpus through separate texts and joins the overall database of words.

2.3 Illustration material of linguistic scientific literature represented in the corpus

The dialectal lexical fund can be represented in the corpus in its complete view in the vast textual massive but the probability of realization of absolutely completed paradigm is very low. We suppose that linguistic "representativeness" of the dialectal corpus must be "enhanced" by the paradigmatic descriptions of the separate dialects (accompanied by illustrations). These are the three factors: a text, a dictionary and data from the scientific literature, which support the creation of truly representative dialectal corpus. Illustration data confirmed in the scientific texts will be also integrated in the textual component of the corpus.

3 Principles of annotation in the corpus

Annotation parameters are selected in accordance with the two requirements set to the corpus by the working group. Those requirements consequently reveal two main tasks of the corpus:

a. Show the most perfect model of the given linguistic area;
b. Show as complete ethno-cultural picture of the given region, as possible.

3.1 Text and meta text characteristics in the corpus are established so, that it becomes possible to study the texts in inter and multi disciplinary ways

We consider especially important to employ these texts with the status of "oral histories". Such precedent has already appeared earlier and on the bases of the texts, obtained for the dialectological purposes, we published two books about two important stages of the Georgian history – about the Muslim population deported out of Meskheti so called Muslim Meskhetians (Beridze, 2005) and about anti-government rebellion in the Georgian Highland, in the period of the II World war (Tsotsanidze, 2004).

Despite of the fact that there exist the fundamental principles of the corpus construction, they can vary according the language peculiarities, specific historic, cultural and linguistic characteristics. The most "variable" are characteristics of the corpus annotation. Especially this is true to the metatextual annotation characteristics. During the XX century, Georgia lived its life as a region of Russia (in the beginning) and then as a Soviet Republic (till 1990). The country could not avoid historical cataclysms that all the countries subordinated to Russia had undergone. Especially important are the moments, which have significantly changed ethnographic, and correspondingly the linguistic situation in Georgia. We mean some waves of migrations – natural migrations from the highlands down to the valleys and compulsory, repressive migrations forced on the population by the communist regime, which nowadays are considered as occurrences of ethnographic trafficking. Recently, the history of such events was "enriched "with the facts of displacement of large groups of population from Abkhazia and South Ossetia. The picture of dialects geographical distribution greatly changed and at the end of the XX c. it differed from that of the beginning. Many "dialectal isles" occurred on the territory of Georgia and the country turned into a "live laboratory" for studying the migration processes and their results. Correspondingly, the parameters of texts description in our corpus are aimed to describe existing complex and diverse situation.

Today, the following characteristics work to describe texts in the corpus:

a) Information about narrator contains several blocks: name, ethnical and regional origin; age, place of birth, occupation; information about parents, their origin (to study speech of the mixed families); information about migration of the narrator or his family: year of migration, which generation was migrated (the

narrator, his father, grandfather or ancestor); type of migration (compulsory, voluntary), kind of dwelling in new places (compact or not compact).

b) Information about the text: who has recorded, who has prepared it for publishing, description of publishing (date, place) and so on;
c) Types of the text: narrative, folklore (prose), folklore (verse), fragments of conversations;
d) Thematic characteristics;
e) Chronotypical description of the text;
f) Form of the text: printed, manuscript (from the old archives), audio, video

3.2 Morphological Tagging (most basic type of linguistic corpus annotation)

As we have mentioned above, the corpus has an ambition to represent one of the most important segment of the Georgian language world – the model of its spatial variation systems. Dialectal material given in the corpusal structure is approached diversely. Some of the corpora are oriented to show interrelations between dialects and the literary language, and thus, to show only specific characteristics of the dialect; other corpora are focused on presentation of dialect as a whole linguistic system and the dialectal data is given not differentiated. We are guided by the second of the named principles, according to which, each form acknowledged in the dialect regardless coincides it with the literary form or not, is referred to as an element of the given dialectal system.

3.3 Primary morphological tagging

Primary morphological tagging implies initial tagging according the morphological groups. On this stage of tagging the whole corpus material is divided into the following clusters: noun, verb, verbal noun, invariant word-forms, non-differentiated stem. On the first stage of morphological tagging we marked non-differentiated stems as a separate group – these are word-forms originated from enclitic of different elements, which only then are described according the all-constituent elements.

Primary morphological tagging process involves: description of the word-form by the primary morphological classificator, its lemmatization, marking literary correspondence of the word-form (the next stage will be tagging by all morphological characteristics – by classificators as well as by word-formation tools).

Introducing literary form in the tagging system will assist in:

Conducting statistic investigation:

- What should be the percentage relation between standard and variation systems inventory in the language sub system so that it could meet minimal requirements to exist as a separate communication model – the requirement of the differentiated language identity;
- At which levels of hierarchy is the standard model inviolability applicable;

- At which levels of hierarchy is the dialectal form more durable against the influence of the literary form (or other dialects);
- At which levels of hierarchy do the "parallel" realizations occur more frequently. To what extent can the future of the language occurrence be predicted with help of parallel realizations.

Conducting search by different characteristics:

- Search by literary form (literary form – dialectal form in 16 sub systems);
- Search by dialectal form (dialectal form – literary form + forms of all sub systems);
- Search by the head form (head form – all realizations of the morphological paradigm;
- Literary head form – all realizations of the morphological paradigm (literary, dialectical and others).

It should be noted here that lemmatization in the corpus happens only on the basis of the acknowledged realization, as in the realism of the Georgian language it is absolutely unacceptable to create the head form on the basis of the morphological model (see Kryuchkova Goldyn 2009). Such categorical approach results from the fact that in Georgian dialects one of the main "sources" of differences is the kaleidoscopic diversity of arrangements of existing morphological models and language inventory. That is why when conducting morphological tagging the head form is represented by:

a. According the realizations illustrated in the text;

b. Verifying this form in the speech of the dialect career;

c. According scientific dialectological literature and dialectal dictionaries;

d. Uncertain and unspecified material should be gathered separately and specified during the future field expeditions.

Certainly, because of some reasons, several of the word forms realizations can not be coincide throughout almost twenty sub systems. However, equaling to the literary head form will allow to reveal "gaps" in description of certain dialects –at the lexical, morphemic or paradigm components levels.

System of tagging at the next stage of "building the corpus" implies the mentioned revealing of gaps and establishing such net and interactive scientific-teaching systems by means of that these gaps can be filed with information. In such way we want to merge the experience in dialectology and linguistic geography in the corpus.

3.4 Possibilities of partially automatic tagging in the corpus

Being oriented on the literary form facilitates solving of two main tasks: partially automatic tagging according the morphological group and correct correlation of the literary language (and other language strata) and the dialectal arsenal.

By means of the morphological analizators established on purpose of the Georgian language studies, according the literary head forms, identification of the dialectal head form can be achieved – lemmatization and automatic annotation on the parts of

speech basis (parts-of-speech-tagging); this surely will entail manual annotation and hard work of deleting occurrences of homonymy.

This technique of lemmatization can not be perfect as it applies only to that part of lexis, the head form of which is confirmed in the text and coincides with the literary form. But refining the process of lemmatization is possible if consider morphophonemic changes characteristic for Georgian dialects and show relative regularity. For example, in the Georgian dialects the stable allomorphs of vowel stems are: 0, -i, semi-vowel –i?, also in the nouns ended in –a, we can encounter a + i > e; correspondingly, if we extend the declension model of nouns with this additional information of vowel stems established for the literary language, it becomes quite realistic not only to identify those vowel stem nouns whose nominative case coincides with the literary form and their dialectal variants as well. "patara" (small) – //patarai//patare…

Similarly, the dialectal form can be identified according the literary word form and this will also enable us to conduct automatic tagging in reverse succession: dialect word-form = literary word-form _ literary head form; e.g. consideration of dialectal variations of the proverbs will provide precise identification of the verb forms, as they differ from the literary correspondences very often only by proverbs. mo-(literary)//me -//ma proverbs identification can result in the identification "ma-itana" "me-iatana" forms with this literary variation "mo-itana" (brought) – and then in relation of this with the literary head form "motana" (to bring). This is sufficient for primary tagging of morphological group and for deep morphological automatic tagging.

In 1986 "The Dictionary of Georgian Morphemes and Modal Elements" was compiled in which dialectal variations of separate morphemes and dialectal variations of the modal elements are registered with great precision. The morphological inventory such as invariant words are represented with limitation as it happens in every language and corresponding dialectal variations are given, so that the list of such words and initial information can be delivered to the database, as a simple identificator. The same can be said about pronouns.

For partial tagging (identification of words with their literary variants) some phonetic rules, evidenced in the dialects with certain regularity, can be employed successfully: e.g. voiced dental consonant "d" can give the voiceless "t" at the end of the word. Consequently, all the words confirmed in the corpus differing only by the final "d" and "t", may be identified as phonetic variations of one form, e.g. adverbial case forms: "kaca-d" (as a man) and "kaca-t", "patarad-patarat" (as a little) and so on. Using such technique for identification enables not only automatic lemmatization but also primary and deep morphological tagging.

One of the main characteristics of the Georgian language is the diversification of the dialects on the morpheme and fundamental levels. Complex and spontanic variations of the vowels and consonants are studied in detail. Using this intellectual resource of the corpus of dialects for automatic annotation is the future goal of our work team.

Summary

The experience has proved that the corpus linguistics more and more extends its functions and purposes and has already gone beyond “utilization” and obtained far more functions, such as ability of interdisciplinary activation, preventive purpose, perspective of developing as a new communication system, etc. The corpus dictates its new functions itself. Very soon the corpora created by us will independently determine the priorities of our work, structure of the material, and so on.

The corpus of the Georgian dialects is under processing now. The main text base still is not completely integrated in the corpus. Besides, we are facing very work - and time - consuming process of the linguistic tagging. The multimedial part (hundreds of hours of audio and video material) has not been included yet in the corpus.

Today we can more surely say that the corpus method of documentation and investigation of the language data is the most effective means of maintaining language and cultural consciousness for a small country like Georgia. Besides that, it differs from the corpora by specially created for the small languages as it is not a static picture, nor a catalogue of grammatical and lexical paradigms, but a ”live” communicative environment, involving the whole language system, language world, with all the characteristics and “genres” of the oral speech. Thus it is not an “enlarged photo” to keep the memory about the language which is on verge of disappearing (Kibrik 2007) but one of the means to prolog its life, its existence.

The corpus is created with the goal of documentation and investigation of the Georgian language world, but it is constructed on the principles, which can be applied to describe any other multi-dialectal and multi-language area.

References

[1] Kibrik A. E., Arkhipov A. V., Daniel M. A., Kodzasov S. V., Meiers, Nakhimovski A. D. (2007). Digital processing of linguistic data for minority languages documentation, *papers of the international conference on computational linguistics, dialogue-2007.*

[2] Ter-Avanesova A. V., Krilov S. A. (2006). Lexical-grammatical databases as a tool of dialectological descriptions, *papers of international conference on computational linguistics, dialogue-2006.*

[3] Kryuchkova O. U., Goldyn V. E. (2008). Textual dialect corpuses a model of traditional rural communication, *papers of the international conference on computational linguistics, dialogue-2008.*

[4] Jorbenadze B., Kobaidze M., Beridze M. (1988) Dictionary of the Georgian morphemes and modal elements, Tbilisi.

[5] Beridze M. (2005). Direct reporting from the past (Meskhetia and Meskhetians, 1918 – 1944), Tbilisi

[6] Ershov A. P. Machinery fund of the Russian language – external setting. *Typed text* http://ershov.iis.nsk.su/archive/eaindex.asp?lang=1&did=11377

[7] Szmercsanyi Benedikt, Hernández Nuria. (2007). Manual of information to accompany the Freiburg Corpus og English Dialects Sampler ("FREDS"), *English department, University of Freiburg.*

[8] Wood Gorodon R. (1969). Dialectology by computer, *International Conference on Computational Linguistics, Sveden.*

[9] McEnery A., Wilson A. (1996). Corpus linguistics.

[10] Sinclair J. M. (1992). The automatic analysis of copora // Directions in corpus linguistics.

[11] Zakharov V.P. (2005). Corpus Linguistics, *Teaching manual.*

[12] Kutuzov A.B. Course of the corpus linguistics, http://tc.utmn.ru/files/corpus_2.pdf

Corpus of Computational Linguistics Texts

Tatiana Bobkova, Mariia Kasianenko, Kuzma Lebedev,
Valentyna Lukashevych, Pavlo Petrenko, and Liubov Grydneva

Computational Linguistics Laboratory, Kyiv National Linguistic University, Ukraine
compling@knlu.kiev.ua

Abstract. The aim of the decision was to compile a corpus of computational linguistics texts and to study its applications in linguistics and lexicography studies. The corpus includes the texts of handbooks, articles in English, Russian and Ukrainian. The whole size of the Corpus is about 500 thousand word-forms in each language. The Corpus is used in linguistics studies, for example statistical research of functioning of words, grammar forms and collocations in scientific texts. A software package was developed to study the texts of the Corpus. The standard principles of text coding compilation were revised. The glossary for searching about thousand computational linguistics terms was designed. This tool includes an explanation of the term, its Russian and Ukrainian translations, and contexts in all sub-corpora.
Keywords: corpus, sub-corpus, text, database, glossary, term.

1 Introduction

The aim of the research described in the article is a development of the principles of compiling a trilingual corpus of computational linguistics texts and studying its possible applications in linguistic researches. Similar corpora have a wide practical use, particularly, in computational terminography, development of systems of automatic text analysis, machine translation and information retrieval systems.

The relevance of this research is in a necessity of systematization and standardization of computational linguistic terminology, functioning in researched professional texts.

The corpus compiled in the laboratory of computational linguistics is trilingual and homogeneous in point of chronology, functional style and text subject.

2 Steps of corpus building

The material of the research is a collection of computational linguistic texts in English, Ukrainian and Russian, the size of which is about 500 thousand word-forms in each language. Compiled in the laboratory of computational linguistics of Kyiv National Linguistic University the corpus is based on the texts of manuals, handbooks and scientific articles of computational linguistics. For instance, the base of the sub-corpus of English texts are the textbooks “The Oxford Handbook of Computational Linguistics” [1], “Speech and Language Processing” [2], and the collected articles

"Computational Linguistics", "Text, Speech and Language Technology", "The Prague Bulletin of Mathematical Linguistics" etc. The Ukrainian sub-corpus is based on the texts of manuals "Computational linguistics" [3] and "Traditional and computational linguistics" [4].

Authentic texts from the end of XX century till nowadays were taken for forming each sub-corpus. The application of the trilingual material of different genres gives an opportunity to include the vocabulary of different linguistic branches (theoretical, structural, applied linguistics) and to represent different schools (American, European, Kyiv, Moscow, St. Petersburg and others). The whole texts of documents are included to the corpus that provides their structural and lexical completeness.

For compiling the corpus an additional program of addressing, which provides texts processing by output data, was developed. Necessary data about a processed text are input into the appropriate fields of the program. The data are the following: the name of the text, surname and initials of the author, the name of the publishing house and the year of edition. An annotation to this text is input separately. Then the program enters information to the appropriate tables in the database. Coding of the texts annotations gives an opportunity in the future to create an automatic system of professional texts abstracting. The scheme of database of the created corpus of computational linguistics texts is stated below.

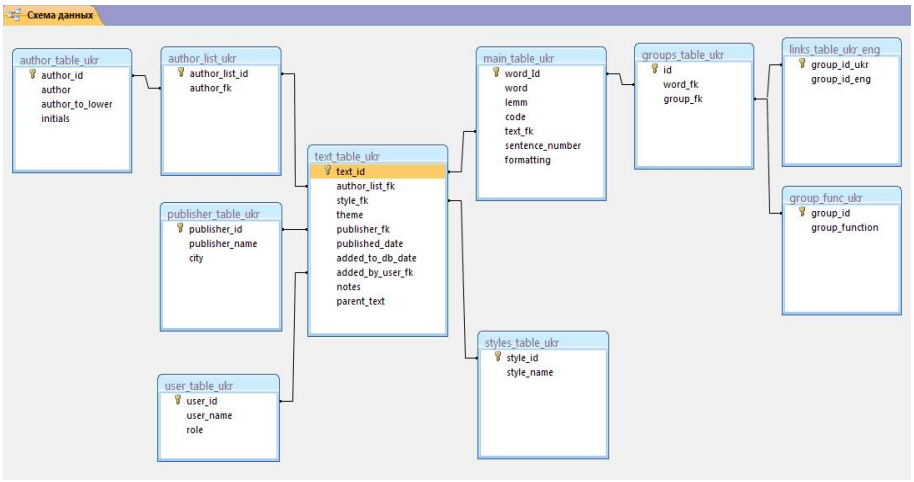


Fig. 1. Structure of corpus database

The tables "**text_table_ukr**" and "**main_table_ukr**" are the most essential ones in the database structure.

The table "**main_table_ukr"** contains texts divided into separate words, their lemmas and grammar codes. Besides, a number of the sentence, where the word occurred, a field with marks about its formatting and an external key-number of the text with it are indicated for each word. The table "**text_table_ukr**" includes

information about texts (a date of publication, a subject, information about the person who added this text to the database, the larger text, a fragment of which is given), and external keys for connection with the tables of the lists of the authors, publishing houses, styles etc.

The tables "**author_list_ukr**", "**author_table_ukr**" provide a treelike structure of data about the authors of the texts that enables to indicate an unlimited amount of authors.

The table "**styles_table_ukr**" contains information about the text style that is input into a separate table for greater convenience.

The table "**publisher_table_ukr**" includes data about all publishing houses.

At this point in the corpus half-automatic morphological coding was carried out, frequency dictionaries of word-forms and a glossary of computational linguistics terms were compiled.

Nowadays a corpus with morphological tagging is the most widespread among other types of corpora. Thus, coding includes not only features of belonging to the part of speech, but also codes of grammar categories specific to this part of speech. The result of the part of speech coding is a text containing monosemantic morphological markers. Thus, in the majority of recent corpora programs available in the Internet are used for the morphological coding.

In the corpus such original programs are used: 1) the program of morphological coding of English verbs on the basis of 15 differential features [5]; 2) the program of word search in the sub-corpus. Words can be searched by word-form, lemma and grammar code. The creation of single system of codes for English-Ukrainian-Russian text corpus is associated with objective difficulties that can be explained, first of all, by different typological characteristics of the analyzed languages. It is known that in the Ukrainian and Russian languages grammar meaning is expressed mainly synthetically, by the means of ending, suffix, prefix, accent change, internal inflexion, suppletive modification. On the contrary in English analytical way of grammar meaning expression prevails (by the means of prepositions, conjunctions, articles, auxiliary verbs, other auxiliary words and word order).

Particularly special attention should be paid to the system of English verb forms that, in comparison with Ukrainian and Russian systems, is characterized by considerable complexity and embranchment. To provide effective work of the corpus correct morphological code should be given to each verb form. The program for automatic morphological coding of verb forms carries out only initial tagging. Results obtained after primary text processing needs manual editing and correction of the mistakes made by the program.

Let's observe the operation principles of the program for morphological coding of English verb in details. For this task such additional programs are used:

1. Program for automatic definition of unambiguous verb forms - the majority of the general list of 526 forms.

2. Program for grammar homonymy of Past Simple and Past Participle verb forms recognition. In this case the user gets a request for manual homonymy clarification.

3. Lemmatization program that assigns lemmas to all verb forms.

The programs mentioned above are based on the processing of the database which includes more than 1500 English verbs with their main forms. During morphological coding the methods of context analysis are used as well.

Described programs considerably facilitate the process of morphological code receiving for the researcher and friendly interface simplifies the process of the correction of the mistakes made during automatic tagging.

3 Glossary of the computational linguistics terms

The final aim of creation of the trilingual corpus of computational linguistics text is the compiling of translation terminological dictionary. That's why the experience of monolingual terminological dictionaries compiling in Ukraine is very useful.

Built by alphabetical principle the terminological dictionary of computational linguistics is an explanatory one. Making up the glossary of terms was done by combining system- and text-oriented approaches.

The register of the glossary was based on the list of English terms with their definitions given in "The Oxford Handbook of Computational Linguistics" [1]. It is explained by the fact that Ukrainian and Russian computational linguistics terminology is mainly supplemented through translation, loan translation or transliteration of English terms.

In the created glossary Ukrainian and Russian English terms matches definitions found in the sub-corpora texts are given as well. Moreover authors refused to conform terms definitions in different languages as it allows to reveal the differences in terms interpretation by different linguistic schools.

To show the real functioning of terms in the glossary the examples of their use in sub-corpora texts are given. Full sentences selected from the analyzed texts are offered as the examples of terms use.

The glossary of English terms of computational linguistics includes about 1000 registered words. The developed software allows to search for English terms, look through its English definition and all the examples of usage in the text massive.

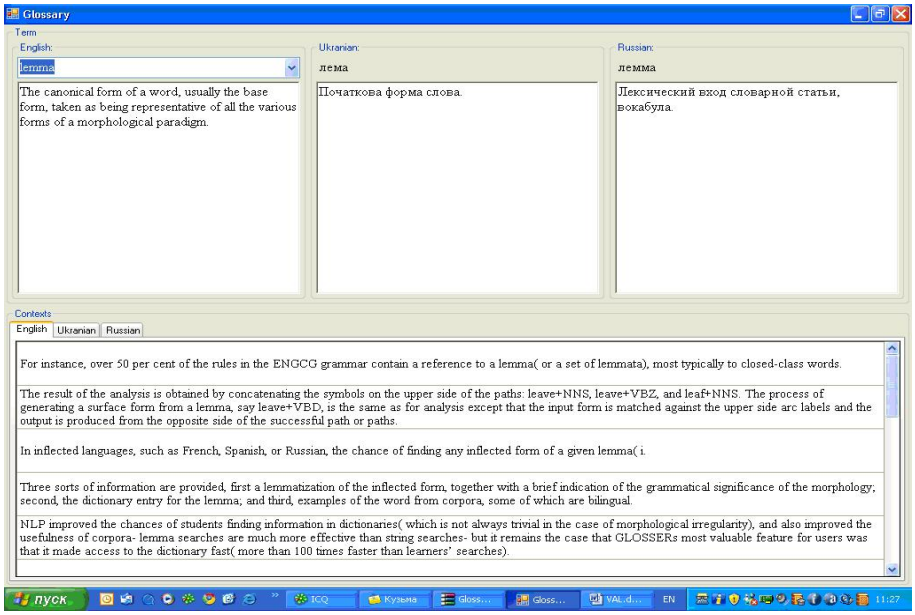


Fig. 2. The example of dictionary entry in the glossary of terms

To view the definition a user may type the term manually or choose it from the list. After pushing the term its definition is given below and the equivalents of the term in Ukrainian and Russian –on the right side. Choosing another working language the user can look through the examples of term equivalent usage in Ukrainian or Russian sub-corpus.

Storing of the glossary as a database allows to add new terms to the formed register. Input and processing of new texts of the corpus automatically increase the quantity of the terms usage examples.

Summary

Designed in the laboratory of computational linguistics of KNLU the trilingual corpus enables future lexicographical and comparatively-typological researches while the term glossary can be used as a reference system with a source database and as an additional module in the machine translation system for professional literature.

References

[1] Mitkov R. (ed.) The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press, 2003.
[2] Jurafsky D., Martin J.H. Speech and Language Processing. Upper Saddle River, N.J.: Prentice Hall, 2000.
[3] Дарчук Н. П. Комп'ютерна лінгвістика. Київ: ВПЦ "Київський університет", 2008.
[4] Перебийніс В. І., Сорокін В. М. Традиційна та комп'ютерна лексикографія. Київ, 2009.
[5] Morphology of English Verb: System and functioning. M.: RGDO, 2008. P. 18-19.
[6] Штерн І.Б. Вибрані топіки та лексикон сучасної лінгвістики. К., 1998.
[7] Дерба С. М. Словник з української термінології прикладної (комп'ютерної) лінгвістики. К., 2007.
[8] Тезаурус з лінгвістичної термінології. http://www.mova.info

"We Only Say We Are Certain When We Are Not": A Corpus-Based Study of Epistemic Stance*

Vaclav Brezina

University of Auckland, New Zealand
vaclav.brezina@gmail.com

Abstract. This paper investigates epistemic stance in speech using two corpora of spoken English: the spoken part of the *BNC* (*British National Corpus*) and *MICASE* (*Michigan Corpus of Academic Spoken English*). In particular, the paper uses corpus evidence to evaluate Halliday's (1994) famous dictum about epistemic stance marking: "The importance of modal features in the grammar of interpersonal exchanges lies in an apparent paradox on which the entire system rests – the fact that we only say we are certain when we are not" (p. 362). The analysis of four epistemic stance markers (*must, certain, sure* and *certainly*) brings evidence in support of Halliday's claim. It appears, however, that this claim is valid only if uncertainty is understood in intersubjective rather than in subjective (psychological) terms.

1 Introduction

In every statement we make, we put across not only some propositional content, but we also indicate how certain or uncertain we are about it (i.e. we indicate our epistemic stance). By indicating our epistemic stance, we simultaneously evaluate the proposition, position ourselves and align with the hearer(s) (Du Bois 2007). Epistemic stance therefore plays a crucial role in any linguistic interaction and spoken interaction in particular.

It is important to realise that epistemic stance can be either explicitly marked (by gestures, intonation or purely linguistic means) or implied by the pragmatics of the speech act. From the perspective of pragmatics, if a speaker makes an unmitigated statement, she commits herself to the truth of the proposition and can be held accountable for it (Searle1985, Holmes 1984a). A simple statement of the type *this is the case* can thus be understood as the neutral way of expressing the speaker's certainty about the propositional content. We can therefore asks the question: why do speakers choose to mark their certainty linguistically forming utterances of the type *this certainly is the case* or *this must be the case*?

A remarkable answer to this question is offered by Halliday (1994), who claims that:

* I would like to thank Dana Gablasova for her invaluable comments on an earlier version of this paper.

> [t]he importance of modal features [i.e. epistemic stance] in the grammar of interpersonal exchanges lies in an apparent paradox on which the entire system rests – the fact that we only say we are certain when we are not (p. 362).

The aim of this paper is to evaluate Halliday's famous dictum using evidence from two corpora of spoken English: the spoken part of the *BNC* and *MICASE*[1]. The former represents the spoken British English from the early 1990s, while the latter captures spoken academic English (American variety) at the turn of the millennia. In short, this paper uses corpus evidence to investigate speakers' choices of expressing certainty in spoken language.

2 Certainty and uncertainty in spoken language

In literature, speaker's certainty and uncertainty about the propositional content of an utterance has been treated under different labels:*modality* (Coates 1987, 2003; Palmer 2001; Nuyts 2001, 2006; DeLancey 2001), *modal evaluation* (Thompson & Hunston 2001), *intensity* (Labov 1984), *evidentiality* (Chafe 1986), *hedging* (Holmes 1984b; Hyland 1996), and *epistemic or epistemological stance* (Barton 1993; Biber et al. 1999; Kärkkäinen 2003; Biber 2006).

Each of these approaches stresses a different aspect of epistemicity. Drawing on the previous research, the present study proposes a comprehensive framework for analysing epistemic stance in spoken language (see Table 1).

Following DeLancey (2001), it can be argued that the function of epistemic stance is to qualify a proposition in relation to *ideal knowledge status.* There are three aspects of this status:

- What is said is known by the speaker by direct experience.
- What is said is assumed to be certainly true.
- What is said is fully consistent with the rest of the speaker's knowledge of the world.

It is important to realise that a unique feature of any epistemic qualification is the fact that unlike speaker's comments on the style of speaking and her evaluative attitude, epistemicity is inherently present in every utterance (although not always linguistically marked). Every statement in language can therefore be said to have the following underlying structure: *epistemic qualification [proposition]*

This underlying structure is reflected in the type of speech act (*pragmatics of epistemicity*) as well as in the linguistic structures which mark the speaker's (un)certainty (*lexico-grammatical epistemicity marking*) – see Table 1. As is apparent from Table 1, linguistic marking of (un)certainty involves different word classes such as adverbs, adjectives, nouns, modal auxiliaries and lexical verbs, which appear in various *epistemic structures* .

[1] *British national corpus* and *Michigan corpus of academic spoken English.*

underlying structure (semantics)		
epistemic qualification [proposition]		
patterns (grammar & pragmatics of epistemic stance)		
patterns		***BNC* examples**
I. (pragmatic aspect of epistemicity) [proposition] unmarked/unmodlised/unmodified illocutionary force		
1. Statement (Searle's assertive speech act)	certainty	S1 [???] alcoholic. S2 [???] ninety nine and a half per cent. S1 [laugh] , erm I'd imagine it's fairly potent. S2 Yes, well **I'm telling you** it is[???]. (explicit, with a *verbum dicendi*) It is eight o'clock though [laugh] (implicit)
2. Question (Searle's directive speech act)	uncertainty	S1: When did you watch Lady And The Tramp? S2: In the lounge. S1: I know where you watched it, **I'm asking you** when you watched it. (explicit, with a *verbum dicendi*) When did you go and get that? (implicit)
II. (lexico-grammatical marking of epistemicity) [proposition] marked		
3. Tag questions		He's got a real little face **hasn't he**?
4. Epistemic matrix clause [epistemic matrix clause][subordinate clause]		**I think that**'s part of the reason why she said you've got your hands full.
		They **seem to** use them a lot, buses, don't they?
		It's pretty **certain that** you can stay.
		It's not **likely to** be over -
		I've come to the **conclusion that** he, I mean, his work is, it was absolutely superb...
5. Epistemic parentheticals, comment clauses, epistemic afterthoughts [main clause] [epistemic parenthetical]		...but they're giving them a packed lunch so **I believe**. ...and the bloke I remembered best of all, **I think**, was oh Cakey
6. Epistemic adverbials [main clause[epistemic adverb]]		They **definitely** had knocked off somebody else..
7. Epistemic modals [main [modal] clause]		They **must** have found it.
III. Combinations of lexico-grammatical markers		
8. Congruous		So **I think probably** that's what it was, but **don't know**.
9. Incongruous		...**I think definitely I think** it **must** be the private scheme...

Table 1. Epistemic structures

Epistemic markers (linguistic signals of epistemic stance) play an important part in the process of *negotiating knowledge*. Knowledge has been traditionally understood as justified true belief (Noddings 2007). Epistemic markers, therefore, first of all, indicate how likely it is that a particular proposition is true. (This is at least one of the functions of these multifunctional items.)

Second, epistemic markers in spoken discourse interaction represent the means of justification of one's beliefs. The justification process, however, does not take place in a vacuum, but in a very complex social setting, which reflects various social power relations. When analysing epistemic patterns, we therefore have to take into consideration the following socio-pragmatic aspects:

- If a speaker claims something she commits herself to the truth of the proposition and can be held accountable for it. In this respect, the degree of certainty can be understood as the degree of commitment to the truth of a particular proposition.
- If a speaker says something with a high degree of certainty, she may find herself in a position of contradicting what another person has said. Hence, this may be potentially face-threatening.
- There exist numerous (often implicit) rules and social norms given by the cultural/social as well as the individual context.

3 Data and methodology

The corpora used in this research (*BNC-spoken part*, *MICASE*) have been chosen to represent general spoken and spoken academic English. From each corpus a sub-corpus of highly interactive language (dialogue) has been extracted so that we can get a better insight into the dynamic process of knowledge negotiation. The details about these sub-corpora (*BNC-CONV* and *MICASE-INT)* can be found in Table 2 below.

Corpora	Tokens	No. of speakers	Speaker's gender	Genre	Discourse mode	Variety of English	Period
BNC-CONV	4,233,955	1525 (?)	33% male 37% female 30% unkn.	informal conversation	highly interactive	UK	early 1990s
MICASE-INT	564,683	520	48% male, 52% female	spoken academic interaction	highly interactive	USA	1998–2001

Table 2. Characteristics of the corpora

BNC-CONV[2], the larger of the corpora, represents the British variety of spoken English from the early 1990s. It contains more than 4.2 million words of transcribed speech (Aston and Burnard 1998, Crowdy 1995).

MICASE-INT, consists of more than half a million words of transcribed academic spoken English (American variety) from various university contexts (lectures, seminars, lab meetings, student presentations, office hours etc.) – see MICASE Manual 2007.

As is clear from the previous discussion (see Table 1 for summary), there are copious lexico-grammatical means, which the speakers can employ to mark certainty in speech. The large epistemic lexico-grammatical structures are, however, relatively difficult to search for in the corpus of spoken language as there exist numerous variants and modifications of these structures (see examples [1] – [4] below).

[1] Certainly has.
[2] They're certainly not quick.
[3] Cos I, I certainly didn't say nothing to her.
[4] Not, oh it was bad, it were it were, well, well stood up all traffic certainly.
(BNC-CONV)

[2] *BNC-CONV* corresponds to the demographic part of the *BNC-spoken* sub-corpus.

Nevertheless, most of these structures contain an epistemic key word (e.g. *certain, conclusion, certainly* etc.) which can be easily searched for using standard concordancing software (e.g. *Xaira*, *MonoConc*). I will call these keywords *epistemic candidates*. An extensive list of epistemic candidates can be found in Biber (2006) and is summarised in Table 3. The epistemic candidates are further divided according to the degree of certainty which they mark and each of the high certainty markers is accompanied by a raw frequency count based on the *BNC-CONV* corpus.

	High certainty (raw frequencies in *BNC-CONV*)	Lower certainty (likelihood)
ADVERBS	actually (3310), always (2813), certainly (426), definitely (503), indeed (207), inevitably (1), in fact (427), never (4113), of course (1217), obviously (672), really (9128), undoubtedly (5), without doubt (0), no doubt (60)	apparently, evidently, in most cases, in most instances, kind of, maybe, perhaps, possibly, predictably, probably, roughly, sort of
ADJECTIVES	apparent (5), certain (288), clear (316), confident (25), convinced (32), correct (97), evident (2), false (55), impossible (43), inevitable (3), obvious (84), positive (33), right (14984), sure (1958), true (726), well-known (0)	doubtful, likely, possible, probable, unlikely
NOUNS	assertion (0), conclusion (21), conviction (3), discovery (5), fact (323), knowledge (40), observation (6), principle (12), realization (1), result (32), statement (34)	assumption, belief, claim, contention, feeling, hypothesis, idea, implication, impression, notion, opinion, possibility, presumption, suggestion
LEXICAL VERBS	conclude (4), demonstrate (11), determine (19), discover (55), find, (2744) know (33526), learn (446), mean (12351), notice (501), observe (7), prove (78), realis/ze (545), recognis/ze (92), remember (2619), see (16527), show (1147), understand (565)	appear, assume, believe, consider, doubt, expect, find, forget, gather, guess, happen, hypothesize, imagine, judge, know, learn, predict, presume, presuppose, pretend, reckon, remember, seem, speculate, suppose, suspect, tend, think
MODALS	necessity must (3014), should (4396), (had) better, have to, got to, ought to (454)	prediction will, would, shall, be going to	possibility can, could, may, might

Table 3. Epistemic candidates according to Biber (2006), Biber et al. (1999)

It is important to bear in mind that not all epistemic candidates function as markers of the speaker's epistemic stance in all contexts, as is evident from examples [5] and [6].

[5] it only acts this way in a certain circumstance... (MICASE-INT)
[6] he was in there and he said my old man knows for certain that at that time they were under fire (BNC-CONV)

In example [5] the adjective *certain* is synonymous with the adjective *particular* and therefore does not contribute to the marking of the speaker's epistemic stance (see 4.2). In example [6], on the other hand, *certain* marks an epistemic stance. However, this does not represent the speaker's epistemic stance, but a reported

epistemic stance. Instances such as [5] and [6] have been excluded from further analysis as irrelevant. In practice, this was done by manually checking all concordance lines with epistemic candidates.

For the purposes of the present study, three epistemic candidates have been chosen: the modal auxiliary *must*, the adjective *certain* and the adverb *certainly*. These represent a variety of lexico-grammatical options available to speakers for marking certainty. Nevertheless, these forms by no means exhaust all the possibilities of marking high degree of epistemic stance (see Table 2). This limitation, therefore, needs to be kept in the back of our mind when we look at and interpret the data.

4 Analysis

4.1 Must

The modal auxiliary *must* is one of the most frequent epistemic candidates. In fact, as previous corpus-based research shows (Holmes 1982, Biber et. al. 1999) modal auxiliaries are the most common linguistic markers of epistemic stance. It is also primarily the epistemic *must* that Halliday has in mind when he makes the claim about the paradox inherent in the epistemic system (see Halliday 1994: 354ff). It is therefore more than justified to start the discussion with this form.

The modal *must* can be either epistemic (as in example [7]), or deontic (as in example [8]). Only the epistemic uses of *must* will, however, be a subject of further analysis as only these are relevant to certainty marking.

[7] Erm it must have been the last day, it must have been Friday... (BNC-CONV)
[8] must try a lot harder. (BNC-CONV)

Corpora	**"must"**[3]		**Epist. stance**				
	AF	NF	AF	NF	%	EV	UN
BNC-CONV	3014	711.9	1929	456	64	195	251
MICASE-INT	91	161.2	54	95.6	59.3	21	9

Table 4. Frequency distribution of must

The data show that marking epistemic stance is the predominant function of *must* both in informal (64 per cent) and academic (59.3 per cent) spoken interaction. This is in accordance with Biber et. al's (1999: 494) finding that face-to-face interaction favours epistemic *must* over deontic *must* since in face-to-face interaction, obligation is expressed more obliquely (i.e. without employing the deontic *must* and thus in a less face-threatening way).

[3] The columns with the heading "must" report the overall frequency of the form *must* (i.e. *must* as an epistemic candidate) in the corpora, whereas the columns with the heading "Epist. stance" show the frequency distribution of *must* as a marker of the speaker's epistemic stance.
AF ... absolute frequency; NF... frequency normalised to the basis of one million; EV... cases of epistemic stance, which are explicitly evidential;UN...cases, in which epistemic *must* occurs in the context of marked uncertainty (i.e. with expressions such as *maybe*, *I think* etc.).

Traditionally (Palmer 1990, 2001), epistemic *must* has been seen as a marker which appears in situations in which the speaker makes judgement on the basis of direct evidence (observation). The corpus data partly confirm this claim showing that indeed epistemic *must* primarily appears in evidential contexts, which are, however, not necessarily based on direct evidence. Cf. example [9], in which the use of *must* is triggered by general experience or "folk" knowledge.

[9] A: I had a very strange dream!
B: Must have been something we'd eaten. (BNC-CONV)

It is interesting to notice that many of these evidential contexts are explicitly evidential (195 examples in *BNC-CONV* and 21 examples in *MICASE-INT*).In theses cases, speakers either provide reasons for their claims (example [10]) or use a deictic reference justifying the claim (example [11])

[10] hm, yeah i, i (don't know) Rasmussen must have been easy or something cuz, it wasn't too hard for, for us (MICASE-INT)
[11] A: Was there an accident? I dunno. This is an ambulance here.
B: An ambulance. so it must be an accident! Look like it doesn't it? (BNC-CONV)

The evidential aspect of the contexts in which *must* appears provides an important clue for evaluating the degree of certainty connected with the use of the epistemic *must.* Although providing reasons for one's claims may strengthen these claims, it at the same time brings about a certain degree of tentativeness. The mere fact that the speaker opts for justification of her statement suggests that in that particular situation the statement is not perceived as obvious and may be challenged.

It is important to bear in mind that it is the pragmatics of the speech act that provides the basic epistemic ground (see section 2). This means that many of our statements are perceived as certain without being linguistically marked for certainty. A simple unmarked statement can therefore be interpreted as unproblematically certain. The speaker's choice to mark the statement using an epistemic marker such as *must* adds another layer of epistemic meaning.

A further insight into this process can be gained by looking at examples, in which an epistemically marked statement appears side by side the same statement, which is unmarked (see examples [12] and [13]).

[12] Yes, it was taken about July that was, must have been. Because I remember... when Tom was on holiday... tha - that ground was all ... dry wasn'tit? (BNC-CONV)
[13] he tipped out a pile of library tickets onto the thing and it was, it was a, it was, must of been, it was hundreds...(BNC-CONV)

In both examples the speaker moves between an unmarked statement and a statement qualified by the epistemic *must.* In example [12], the speaker first makes a statement*it was taken about July,* which itself is rather tentative (since the speaker employs the approximator *about*). What follows is a process of justification (in this

case taking place within a single turn), in whichthe speaker provides arguments for her original statement. [13] is then an example of how a speaker moves back and forth between the unmarked and the marked statement in the process of thinking. These examples suggest that epistemic *must* is employed in situations in which the speaker's statement is far from being unproblematically certain.

Moreover, the data indicatethat epistemic *must* occurs relatively frequently (251 examples in *BNC-CONV* and 9 examples in *MICASE-INT*) in the context of explicitly marked uncertainty as in examples [14] and [15].

[14] He must have gone out. I dunno.. (BNC-CONV)
[15] ...i mean i think that's something that must be sort of, maybe the U-S gets of that to some degree... (MICASE-INT)

4.2 Certain (and sure)

The adjective *certain* is another epistemic candidate which will be discussed in relation to Halliday's claim. In fact, the structure *I'm certain (that)...* is the most explicit expression of speaker's high degree of certainty about a proposition. It closely reflects the underlying epistemic pattern (see section 2) and there is a clear attribution of the epistemic stance to the speaker.

There are however, two important issues which need to be considered in relation to this epistemic candidate. First, there are two basic uses of the adjective *certain,* which are reflected in the corpora and which are found in most English dictionaries (e.g. *COBUILD, CALD, OALD, OED*[4]). *Certain* can indicate either speaker's certainty about something (epistemic stance) as in [16] or it can be used to refer to a particular thing without specifying this item as in example [17].

[16] Oh I'm certain it is. (BNC-CONV)
[17] i think ideally they wanted you to write poems on a certain theme and put 'em into a book... (MICASE-INT)

It is interesting to notice that both in ordinary conversation as well as in academic interactions the latter use of the form *certain* (i.e. *certain* meaning *particular*) is much more frequent than the former use (i.e. epistemic stance) -see Table 5. In fact, there is no attestation of the adjective *certain* expressing the speaker's epistemic stance in *MICASE-INT* and only 17 instances of it in *BNC-CONV*. How can this fact be accounted for? This question leads us to the second consideration.

[4] *Collins COBUILD Advanced Learner's English Dictionary, Cambridge Advanced Learner's Dictionary, Oxford Advanced Learner's Dictionary, Oxford English dictionary.*

Corpora	"certain"[5]		Epist. stance		
	AF	NF	AF	NF	%
BNC-CONV	288	68	17	4	6
MICASE-INT	124	219.6	0	0	0

Table 5. Frequency distribution of certain

The data show that speakers are rarely as explicit as to use the structure *I'm certain (that)...* Nevertheless, speakers have a choice to use a similar structure *I'm sure (that)...,* which is less formal.As is clear from Table 3, the adjective *sure* is considerably more frequent in casual conversation (*BNC-CONV*) than the adjective *certain*. The phrase *I'm sure* appears 563 times (133 per million words) in the *BNC-CONV* corpus and 45 times (80 per million words) in *MICASE-INT*.

There are two major functions of the phrase *I am sure (that)...* First, in informal conversation as well as in the academic interactions, the phrase is often connected with expressing encouragement as in example [18]. Here the social function is more dominant than the epistemic one.

[18] A: Well, the first one's [job] always gonna be the worst one!
B:Mm! Certainly is! But I say, **I'm sure** your capable of taking the bull by the horns. (*BNC-CONV*)

Second, the phrase appears as a linguistic marker of epistemic stance proper. When *I'm sure* is employed with this function, this is very often in situations in which the certainty of a statement is in dispute: *I'm sure* is thus often contrasted with *I'm not sure* or *I don't know* as in examples [19] and [20]. It may also appear in situations, in which the certainty of a claim is challenged by the addressee (example [21]).

[19] A:...it's either Iceflow or Iceland **I'm not sure** which one it is
B: **I'm sure** it was Iceland but that in King Street is Iceflow ***in't it***? (*BNC-CONV*)

[20] agree also because of the fact that children will see this, and **i don't know i'm sure** parents wouldn't want them to see this or for them to emulate this behavior (*BNC-CONV*)

[21] A: she's wrong.
B:**Mm?**
A: **I'm sure** Richard said Isle of Man. She keeps saying, Isle of Wight Isle of Wight.. (*BNC-CONV*)

[5] The columns with the heading "certain" report the overall frequency of the form *certain* (i.e. *certain* as an epistemic candidate) in the corpora, whereas the columns with the heading "Epist. stance" show the frequency distribution of *certain* as a marker of the speaker's epistemic stance.
AF ... absolute frequency; NF... frequency normalised to the basis of one million.

4.3 Certainly

The epistemic adverb certainly, as Simon-Vandenbergen and Aijmer (2007) point out, can be considered "a prototypical adverb in the field of modal certainty" (p. 85). Although it is not as common as the adverbial epistemic candidates of high certainty degree such as *of course, really, actually* or *always* (see Table 3), it comes immediately to one's mind, when one starts considering the various ways in which we can say that we are certain.

This does not mean that *certainly* is used exclusively to express the speaker's epistemic stance. In spoken discourse, it is sometimes used as a positive response to a request or a question as in [22] and [23].

[22] A: Please may I have some tea?
B: **Certainly** my dear. (*BNC-CONV*)

[23] A: can I leave it open?
B: Mhm, **certainly**. (*BNC-CONV*)

Nevertheless, as is clear from Table 6, the predominant function of *certainly* is to mark the speaker's epistemic stance (89 per cent of cases in *BNC-CONV* and 100 per cent of cases in *MICASE-INT*). The question to ask at this point is: What are the typical contexts in which *certainly* appears?

Corpora	**"certainly"**[6]		**Epist. stance**				
	AF	NF	AF	NF	%	CON	UN
BNC-CONV	426	100.6	379	89.5	89	79	46
MICASE-INT	49	86.8	49	86.8	100	14	11

Table 6. Frequency distribution of certainly

The data indicate that *certainly* often appears in a combination with an epistemic marker with a lower degree of certainty such as *I think, almost* or a tag question (see examples [24] and [25]).

[24] um, **i think** that's, **certainly**, can be useful and ... (*MICASE-INT*)
[25] They're **certainly** not as old as ours **are they**? (*BNC-CONV*)

In these cases, the speaker undermines a strong epistemic claim by a lower certainty marker. This can be explained by reference to the dynamics of the spoken interaction, in which the speaker often seeks an approval from the addressee in the process of knowledge negotiation.The incongruous combination of two epistemic markers thus leaves more space for the addressee to disagree with the speaker

[6] The columns with the heading "certainly" report the overall frequency of the form *certainly* (i.e. *certainly* as an epistemic candidate) in the corpora, whereas the columns with the heading "Epist. stance" show the frequency distribution of *certainly* as a marker of the speaker's epistemic stance. AF ... absolute frequency; NF... frequency normalised to the basis of one million; CON...the contrastive use of *certainly*;UN...cases, in which epistemic *certainly* occurs in the context of marked uncertainty (i.e. with expressions such as *I think* etc.).

without threatening the speaker's face. These combinations can therefore be said to express a certain kind of socially motivated uncertainty.

However, even if we look solely at instances, in which *certainly* appears as the only epistemic marker, we can very often find this adverb in the context of doubt (79 examples in *BNC-CONV* and 14 examples in *MICASE-INT*). As Simon-Vandenbergen and Aijmer (2007: 96) point out, "*certainly* often functions in a context of contrast with uncertainty." This can be shown on the following examples:

[26] related to the weather or not i don't know. but **certainly** with it being darker when the when the the you know the cloud... (*MICASE-INT*)

[27] Anyway, call it south London. It's **certainly** not a north London accent. (*BNC-CONV*)

[28] she was buying a house for whatever it was it **certainly** was a lot of money. (*BNC-CONV*)

[29] **Certainly** not in my age group. (*BNC-CONV*)

In all the examples above, the speakers use *certainly* in a situation, in which they lack the relevant precise knowledge. In these contexts, *certainly* introduces a piece of information, which is usually vague or not as relevant, but speakers can be relatively certain about it (in contrast to the piece of information which they do not know). The prototypical structure in these cases is *I don't know this, but certainly....*

5 Conclusion

We have reached a point at which we can evaluate Halliday's claim about the markers of certainty in English. Three general points which emerged from the data are worth emphasising at this stage.

First, all of the certainty markers in this study (*must, certain/sure* and *certainly*) play an important role in the process of knowledge negotiation. In the context of spoken face-to-face interaction (be it informal conversation or academic interaction), any piece of information which may not be obvious either to the speaker or the hearer is likely to be epistemically marked. Although in a situation like this, one would expect an epistemic marker of lower degree of certainty (such as *maybe* or *possibly*) to be used, it is often the case that *must, certain/sure* and *certainly* appear in this context.

Second, these epistemic markers are closely connected with uncertainty. The uncertainty may be present in the immediate context as in the incongruous combinations such as *I think it must, I don't know I'm sure, I think it certainly* etc. or it can be implied by the need of justification of the speaker's statements (the evidentiality context of *must*). Sometimes also (as is often the case with the adverb *certainly*) the uncertainty about one piece of information is contrasted with certainty

about another piece of information which, however, is not as precise or relevant the original piece of information.

Third, the notion of uncertainty which appropriately captures the dynamics of the spoken interactions is not the subjective (psychological) uncertainty, but rather intersubjective uncertainty, which reflects the process of knowledge negotiation and knowledge co-construction in dialogue. As linguists, we do not claim to have a direct access to the processes that take place in the speaker's mind. For this reason, we cannot say much about the speaker's subjective uncertainty. However, thanks to corpora, we do have an access to the typical epistemic patterns, in which, as the data indicate, markers of certainty and uncertainty often occur side by side.

To conclude, we can say that Halliday's claim is to a large extent justified, if uncertainty is understood as intersubjective uncertainty. In this light, Halliday's claim can be paraphrased as: *We only say we are certain when there is a need to negotiate the validity of what we say.*

References

Aston, Guy, and Lou Burnard. 1998. *The BNC handbook: Exploring the British national corpus with Sara*. Edinburgh: Edinburgh University Press. Barton, Ellen L. 1993. Evidentials, argumentation, and epistemological stance. *College English* 55 (7):745-769.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Chafe, Wallace. 1986. Evidentiality in English conversation and academic writing. In *Evidentiality: The linguistic coding of epistemology*, edited by W. Chafe and J. Nichols. Norwood: Ablex.

Coates, Jennifer. 1987. Epistemic modality and spoken discourse. *Transactions of the Philological society* 85 (1):110-131. 2003. The role of epistemic modality in women's talk. In *Modality in contemporary English*, edited by R. Facchinetti, M. Krug and F. Palmer. Berlin: Walter de Gruyter.

Crowdy, Steve. 1995. The BNC spoken corpus. In *Spoken English on Computer*, edited by G. Leech, G. Myers and J. Thomas. Harlow: Longman.

DeLancey, Scott. 2001. The mirative and evidentiality. *Journal of pragmatics* 33 (3):369-382.

Du Bois, John W. 2007. The stance triangle. In *Stancetaking in discourse: subjectivity, evaluation, interaction*, edited by R. Englebretson. Philadelphia: John Benjamins.

Halliday, M. A. K. 1994. *An Introduction to functional grammar* London: Edward Arnold.

Holmes, Janet. 1982. Expressing Doubt and Certainty in English. *RELC Journal* 13 (2):9-28. 1984a. Modifying illocutionary force. *Journal of Pragmatics* 8 (3):345-365. 1984b. Hedging your bets and sitting on the fence: Some evidence for hedges as support structures. *Te Reo* 27 (1):47-62.

Hyland, Ken. 1996. Talking to the academy: Forms of hedging in science research articles. *Written communication* 13 (2):251-281.

Kärkkäinen, Elise. 2003. *Epistemic stance in English conversation: A description of its interactional functions, with a focus on I think*: John Benjamins. 2007. The role of *I guess* in conversational stancetaking. In *Stancetaking in discourse: subjectivity, evaluation, interaction*, edited by R. Englebretson. Philadelphia: John Benjamins.

Labov, William. 1984. Intensity. In *Meaning, form, and use in context: Linguistic applications*, edited by D. Schiffrin. Washington, D.C: Georgetown University Press.

MICASE Manual: The Michigan Corpus of Spoken Academic English. 2007. Place Published: English Language Institute, The University of Michigan. `http://lw.lsa.umich.edu/eli/micase/MICASE_MANUAL.pdf` (accessed July 2009).

Noddings, Nel. 2007. *Philosophy of education*. Boulder: Westview Press.

Nuyts, Jan. 2001. *Epistemic modality, language, and conceptualization: A cognitive-pragmatic perspective*. Amsterdam: John Benjamins. 2006. Modality: overview and linguistic issues. In *The expression of modality*, edited by W. Frawley. Berlin: Mouton de Gruyter.

Palmer, Frank Robert. 2001. *Mood and modality*. Cambridge: Cambridge University Press. 1990. *Modality and the English modals*. London: Longman.

Searle, John R. 1985. *Expression and meaning: Studies in the theory of speech acts*. Cambridge: Cambridge University Press.

Simon-Vandenbergen, Anne-Marie, and Karin Aijmer. 2007. *The Semantic field of modal certainty. A corpus-based study of English adverbs*. Berlin: Mouton de Gruyter.

Thompson, Geoff, and Susan Hunston. 2000. Evaluation: An introduction. In *Evaluation in text: Authorial stance and the construction of discourse*, edited by S. Hunston and G. Thompson. Oxford: Oxford University Press.

A Model for Corpus-Driven Exploration and Presentation of Multi-Word Expressions

Annelen Brunner and Kathrin Steyer

Institute for the German Language, Mannheim, Germany

Abstract. In this paper we outline our corpus-driven approach on detecting, describing and presenting multi-word expressions (MWE). We make use of large corpora and statistical data to explore and visualize the rich interrelations, patterns and types of variances of MWE. In order to do this, we have developed a method of linguistically interpreting collocational data in a structured way (cf. [Steyer/Brunner 2009]). Several levels of abstraction build on each other: surface patterns, Lexical realizations (LR), MWE and MWE patterns. Generalizations are made in a controlled way and in adherence to corpus evidence. The method helps to identify and describe MWE in a way that gives credit to their flexible nature and their role in language use.

1 Methodological foundations

Our approach is corpus-driven as defined by Tognini-Bonelli who states:

> "In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence." [Tognini-Bonelli 2001, p. 84]

Following this basic principle, we work empirically with large quantities of corpus data and generate our linguistic hypotheses and generalizations bottom up. The following steps are crucial to our interpretative practice (cf. [Steyer/Lauer 2007, p. 494]):

- Study of all evidence of the corpus and acceptance of this evidence: We use collocation profiles as well as pattern matching to get a starting point for our analysis that is as close to real life usage of language and as objective as possible.
- Generation of hypotheses on the basis of the evidence: We take interactive steps in formulating and refining pattern matching queries to study the evidence.
- Empirical checking of those hypotheses: We check the results of our queries for plausibility and revise if necessary.
- Generalization leads to usage rules: In our model, generalization happens on several hierarchical levels and is detailed by narrative comments if necessary. Usage is always the key factor for justifying generalization.

We have a broad concept of MWE, which is heavily influenced by experience with empirical language data and centers around usage. Our German label – Usuelle Wortverbindungen (first used in [Steyer 2000]) – can be translated as 'MWE which are common in usage'. We adhere to Firth's contextual theory of meaning, here summarised by Tognini-Bonelli:

> "In the Firthian framework the typical cannot be severed from actual usage, and 'repeated events' are the central evidence of what people do, how language functions and what language is about." [Tognini-Bonelli 2001, p. 89]

Following this idea, we regard MWE as conventionalized patterns of language use that manifest themselves in recurrent syntagmatic structures and have acquired a distinct function in communication (cf. [Feilke 2004]; cf. [Brunner/Steyer 2007], [Steyer/Brunner 2009]). They must have at least two concrete lexical components, but may also contain abstract components representing a certain subset of lexical items or even a general grammatical class. Neither idiomaticity nor idiosyncrasy on a grammatical or lexical level is a necessary criterion for MWE in our definition. MWE can have a perfectly regular structure, as long as they work as functional chunks in language use.

Our approach to analysis is similar to that of Hanks detailed in the description of his Corpus Pattern Analysis (CPA):

> "Concordance lines are grouped into semantically motivated syntagmatic patterns. Associating a 'meaning' with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns. The identification of a syntagmatic pattern is not an automatic procedure: it calls for a great deal of lexicographic art. Among the most difficult of all lexicographic decisions is the selection of an appropriate level of generalization on the basis of which senses are to be distinguished." [CPA]

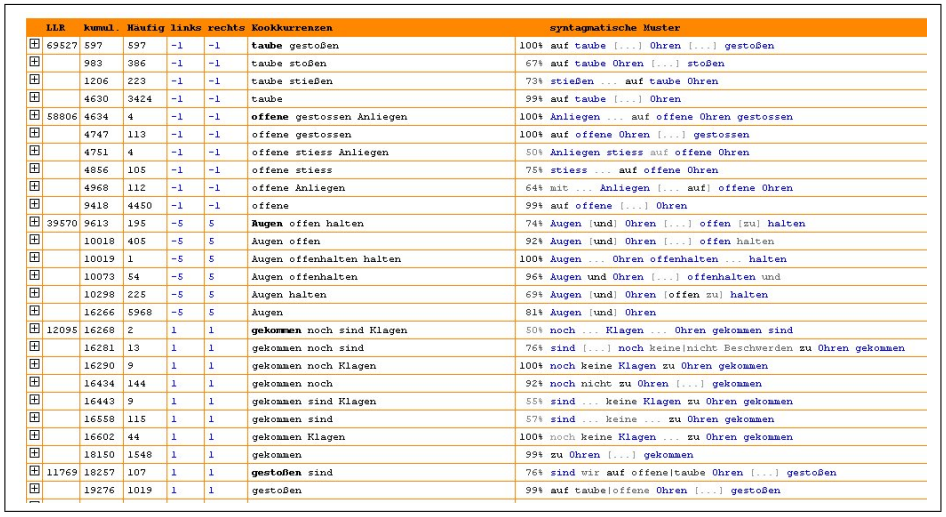
CPA aims at describing single words (cf. also [Hanks 2008]), while we are interested in MWE, which adds an additional level of complexity as identifying the surface form itself requires an interpretative effort. To handle the difficulties of generalization, our model has several hierarchical levels, which will be presented below.

2 Model of analysis

2.1 Prerequisites

The basis of our work is collocation profiles, computed from a very large corpus of written German, DeReKo (Deutsches Referenzkorpus), which currently consists of over three billion tokens (cf. [KLa]). The sophisticated method used for generating these profiles was developed by Cyril Belica ([Belica 1995]). It takes a target word form as input and computes the word forms that appear in the vicinity of this target word more often than statistically expected. These partner word forms are clustered on multiple levels. The KWIC (key word in context) lines from the corpus are grouped into collocation clusters, according to the word forms they contain. The total of all clusters generated for a target word form is called its collocation profile. Collocation analysis is available for the IDS corpora via the COSMAS II corpus research tool (`http://www.ids-mannheim.de/cosmas2`) and can be customized in various ways. For details see [KLb].

Starting point for our study of MWE is the collocation profile of a target word form. Figure 1 shows a snapshot from such a profile for the word form *Ohren [ears]*. Though collocation analysis offers lemmatization, we do not use this setting, neither



LLR	kumul.	Häufig	links	rechts	Kookkurrenzen	syntagmatische Muster
69527	597	597	-1	-1	**taube** gestoßen	100% auf taube [...] Ohren [...] gestoßen
	983	386	-1	-1	taube stoßen	67% auf taube Ohren [...] stoßen
	1206	223	-1	-1	taube stießen	73% stießen ... auf taube Ohren
	4630	3424	-1	-1	taube	99% auf taube [...] Ohren
58806	4634	4	-1	-1	**offene** gestossen Anliegen	100% Anliegen ... auf offene Ohren gestossen
	4747	113	-1	-1	offene gestossen	100% auf offene Ohren [...] gestossen
	4751	4	-1	-1	offene stiess Anliegen	50% Anliegen stiess auf offene Ohren
	4856	105	-1	-1	offene stiess	75% stiess ... auf offene Ohren
	4968	112	-1	-1	offene Anliegen	64% mit ... Anliegen [... auf] offene Ohren
	9418	4450	-1	-1	offene	99% auf offene [...] Ohren
39570	9613	195	-5	5	**Augen** offen halten	74% Augen [und] Ohren [...] offen [zu] halten
	10018	405	-5	5	Augen offen	92% Augen [und] Ohren [...] offen halten
	10019	1	-5	5	Augen offenhalten halten	100% Augen ... Ohren offenhalten ... halten
	10073	54	-5	5	Augen offenhalten	96% Augen und Ohren [...] offenhalten und
	10298	225	-5	5	Augen halten	69% Augen [und] Ohren [offen zu] halten
	16266	5968	-5	5	Augen	81% Augen [und] Ohren
12095	16268	2	1	1	**gekommen** noch sind Klagen	50% noch ... Klagen ... Ohren gekommen sind
	16281	13	1	1	gekommen noch sind	76% sind [...] noch keine|nicht Beschwerden zu Ohren gekommen
	16290	9	1	1	gekommen noch Klagen	100% noch keine Klagen zu Ohren gekommen
	16434	144	1	1	gekommen noch	92% noch nicht zu Ohren [...] gekommen
	16443	9	1	1	gekommen sind Klagen	55% sind ... keine Klagen zu Ohren gekommen
	16558	115	1	1	gekommen sind	57% sind ... keine ... zu Ohren gekommen
	16602	44	1	1	gekommen Klagen	100% noch keine Klagen ... zu Ohren gekommen
	18150	1548	1	1	gekommen	99% zu Ohren [...] gekommen
11769	18257	107	1	1	**gestoßen** sind	76% sind wir auf offene|taube Ohren [...] gestoßen
	19276	1019	1	1	gestoßen	99% auf taube|offene Ohren [...] gestoßen

Fig. 1. Collocation profile of *Ohren* (computed 29 July 2009 via COSMAS web)

for the target word form nor when computing its collocates. This is because empirical research shows that the behavior and contexts of different realizations of a lemma are often quite different. These distinctions would be obfuscated by lemmatization. Also, lemmatization of a word form is already an abstraction and we want to be very careful not to make assumptions. In this respect, we follow Sinclair who pointed out:

> "There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity." [Sinclair 1991, p. 8]

Consequently, we study the profiles of several word forms which belong to the same lemma separately and make generalizations only at a much later stage of analysis.

We thus start with a collection of KWIC lines which contain the non-lemmatized target word form (e.g. *Ohren [ears]*) and are grouped according to the collocates that have been identified for this target word. This gives us a very good starting point, as the statistical method has already detected regularities within the data with extremely few a priori assumptions. Our goal now is to make use of the information given by corpus data and statistical analysis in a structured and controlled way. We have designed four different levels of abstraction:

- The level of surface patterns
- The level of Lexical realizations (LR)
- The level of MWE
- The level of MWE patterns

Each of these levels builds on the previous and on each level we work manually when grouping, correlating and commenting on the phenomena we observe. The model is

thus based on automatically pre-structured data, but is itself a strategy of controlled human interpretation.

As a main example in this paper, we will look at the MWE *Musik in den Ohren*, literally translated as *music in the ears*, which is similar to the English MWE *music to the ears*. *Musik* is a significant collocation partner of *Ohren* and the KWIC lines containing these two word forms in an appropriate distance are grouped in a collocation cluster. This cluster is the starting point for our analysis. Table 1 shows an excerpt of the relevant KWIC lines.

M06	"klingt wie Musik in meinen	Ohren",	sagte die Sozialbürgermeisterin und
MLD	"Das ist Musik für unsere	Ohren",	sagt Ursula Schmitz. Musik, von
S94	Die Musik kann für westliche	Ohren	eine Qual sein; sie ist
O94	Chancen.. ." Musik in unseren	Ohren.	Doch auch wenn uns wohltut,
O94	Uhren,	Ohren,	Musik In der Grazer ESC
O94	wahre Musik in Horst Fendrichs	Ohren.	"Es leben die PS.. ."
O94	das war Musik in den	Ohren	des Akustik-Sachverständigen. Auch der
O95	Motoren wieder Musik in seinen	Ohren.	Die ersten vier
O95	Sonderling, Künstler, der "die	Ohren	voll Musik hat und den
O95	sind auch Musik in den	Ohren	des Liebespartners. Was gehört noch
O95	Lust sind Musik in den	Ohren	des Liebespartners
O96	ist wahre Musik in seinen	Ohren:	Seine Liebe zur Eisenbahn ließ
O96	Gästen. Musik in Carsten Kelms	Ohren,	denn wenn's bei denen im
O96	narrative Musik halten Augen und	Ohren	immer wieder lustvoll auf Trab.
O97	- zeitgenössische Musik für junge	Ohren	(Ossiach, 13. bis 16. 7.).
O97	Während es etwa in den	Ohren	vieler wie Musik klingt, wenn
O97	Musik in des versöhnten Ombudsmans	Ohren!	Die Starparade ist der Höhepunkt
O97	euch" wie Musik in den	Ohren.	Aber die Papst-Visite wurde ebenso
O97	angekündigt. Musik in meinen	Ohren.	
O99	der lauten Musik kurz die	Ohren	zuhielt... Gefeiert wird in St.

Table 1. KWIC lines from the collocation cluster *Ohren – Musik*

2.2 The level of surface patterns

On this level the KWIC lines which have been grouped by collocation analysis are subjected to further structuring. For this we use a query syntax based on regular expressions. The queries are used to identify and group lines with a similar syntagmatic structure which serve then as a basis for the analytic steps that follow.

This step is necessary, as our definition of MWE calls for a common syntagmatic structure of the instances of an MWE while collocation analysis looks at word form surfaces without considering the syntactic connection between them. Thus, it sometimes sorts instances of different MWE which share the same lexical material into a single cluster or assigns instances of the same MWE which have different lexical material (e.g. because of orthographical variance or different realizations of the same lemma) to different clusters. Humans, as opposed to the computer, can decide which surface similarities are important for the task of identifying and describing MWE and formulate surface patterns designed to gather the correct instances.

Pattern matching is also a valuable asset when exploring the variability of an MWE. The patterns can be formulated more or less restrictively and the researcher can observe how many KWIC lines – i. e. instances of realization of the MWE in the corpus –

are captured. It is also possible to define gaps in the patterns and study the fillers for these gaps. This gives a very good indication what features are really important for the structure of a MWE, how it can be modified and what is its core meaning.

The definition of surface patterns is an iterative process. Often the form of the patterns has to be adapted when observing the results of the previous try. This process reflects an ever deepening understanding of what is relevant in the MWE structure.

The cluster *Ohren – Musik* in Table 1 above is a good example for a collocation cluster which contains instances of different MWE, for example *westliche Ohren [western ears]* or *die Ohren zuhalten [to cover the ears]*. Here are some examples of search patterns which are used to filter out instances which are relevant for the description of the MWE *Musik in den Ohren*. (`#*` stands for an arbitrary number of unspecified word forms; `N(den)` stands for "not *den*"; `ist|war` stands for "*ist* or *war*".)

(1) `Musik in #* Ohren`
(2) `Musik in den Ohren`
(3) `Musik in #* N(den) Ohren`
(4) `Das ist|war Musik in #* Ohren`

You can see that the search patterns differ in their restrictiveness and that, though for sake of simplicity the 'name' of the MWE was given above as *Musik in den Ohren*, not all realizations actually take exactly this form.

Search pattern 3 is specifically designed to capture all realizations that do *not* use the definite determiner *den*. We can now examine the hits of this search pattern and learn from the filler list for the gap that possessive pronouns and genitive forms referring to persons are also common when this MWE is realized. However, frequency counts show that the realization with *den* appears almost three times as often in our corpus.

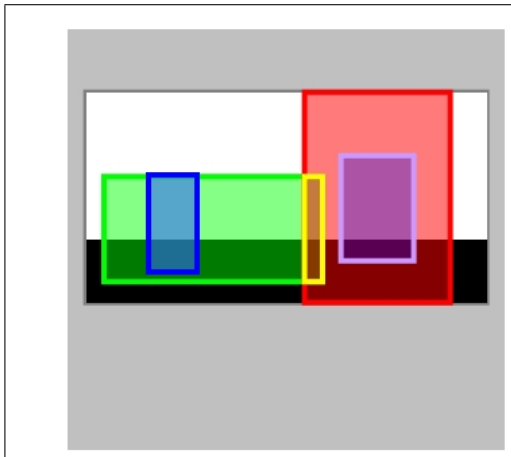
Search pattern 4 is an example of a very specific pattern. It captures a common way of using the MWE which is very stable, though it covers only a relatively small section of all instances of the MWE in our corpus.

2.3 The level of Lexical realizations

The level of Lexical realizations (LR) is a step between the surface patterns and the actual MWE. LR represent typical realizations of an MWE in the corpus. This intermediate level of analysis has been introduced to account for the fact that MWE are very flexible and subject to much variation. When generalizing immediately to a single typical form, many of these nuances would be lost. LR focus on different kinds of realization of the same MWE and offer a chance to comment on them.

From empirical experience, we have defined several types of LR. This typology is quite general and reflects very basic mechanisms of language. The types will be presented below and exemplified by the LR of the MWE *Musik in den Ohren*. Figure 2 illustrates how the KWIC lines from the collocation cluster are assigned to several LR (each subsuming the hits of one or more search patterns) and how the LR tree for this MWE is built up.

Core LR We assume that there is a core structure which is necessary for the MWE to be recognizable. This structure is captured in the Core LR. Often, the structure of the



KWIC lines of the profile
Ohren – Musik,
as captured by the search patterns

LR group: Musik in X Ohren [music in X ears]
Search pattern: Musik in #* Ohren

Core LR: Musik in den Ohren [music in the ears]
Search pattern: Musik in den Ohren

Core Variant LR: Musik in N(den) Ohren [music in N(the) ears]
Search pattern: Musik in N(den) Ohren

Extension LR: Musik in X Ohren sein [to be music in X ears]
Search pattern: Musik in #* Ohren ist|sind|war|waren|gewesen
Search pattern: ist|sind|war|waren|gewesen Musik in #* Ohren

Extension LR: Das ist/war Musik in X Ohren [That is/was music in X ears]
Search pattern: Das|das ist|war Musik in #* Ohren

Extension LR: wie Musik in X Ohren sein [to be like music in X ears]
Search pattern: Wie|wie Musik in #* Ohren #* ist|sind|war|waren|gewesen
Search pattern: ist|sind|war|waren|gewesen #* wie Musik in #* Ohren

Extension LR: wie Musik in X Ohren [like music in X ears]
Search pattern: Wie|wie Musik in #* Ohren

Extension LR: wie Musik in den Ohren klingen [sound like Musik in the ears]
Search pattern:
klingen|klingt|klinge|klang|klangen|geklungen wie Musik in #* Ohren
Search pattern:
Wie|wie Musik in #* Ohren klingen|klingt|klinge|klang|klangen|geklungen

LR tree of the MWE *Musik in den Ohren*

Fig. 2. LR structure of the MWE *Musik in den Ohren* in relation to the KWIC lines of the cluster *Ohren – Musik*

Core LR is also the most general and subsumes the largest number of KWIC lines, i.e. instances of the MWE in the corpus.

Example: The Core LR *Musik in den Ohren* subsumes more than half of all the instances of the MWE and can thus be regarded as the most common realization – especially as the surface forms of alternative realizations are not as stable.

Core Variant LR The core can be subject to variations on the surface, which are documented in the Core Variant LR. The Core Variant LR is defined relative to the Core LR and differs from its structure in at least one respect.

Example: The Core Variant LR *Musik in N(den) Ohren* subsumes all cases in which the determiner *den* is not present. The fillers for the search pattern gap `N(den)` are presented prominently in form of a filler list and it becomes clear that especially possessive pronouns or genitive forms referring to persons take the place of the determiner. It is justified to differentiate between a Core LR and a Core Variant LR instead of just defining a more general Core LR with the search pattern `Musik in #* Ohren`, because it allows us to highlight the fact that these two different types of realization exist and to show their relative frequencies and nuances in meaning.

Extension LR The structure of the Core LR can be extended by additional elements, which are not mandatory for the structure of the MWE, but are still frequent and typical for the way the MWE is used. There can be internal extensions, which appear between the elements of the core structure, or external extensions, which are added before or after the core. Extensions have to be connected to the core syntactically. They are for example verbal constructions, modifiers, object extensions or prepositional phrases.

Example: There are several Extension LR in the LR tree for the MWE *Musik in den Ohren*. All of them extend the structure captured by the Core LR and Core Variant LR in different ways. As you can see in Figure 2, some of them demonstrate that an LR can subsume more than one search pattern. Usually search patterns that only account for word order or grammatical variances are bundled together, unless one of them is extremely prominent or idiosyncratic.

On the first level, there are the LR *wie Musik in X Ohren* and the LR *Musik in X Ohren sein*. Those LR have in common that they add one element to the core structure. Note that they both capture instances of the Core LR structure as well as of the Core Variant LR structure – the difference between those structures has been highlighted already, so it is not necessary to separate them in the Extension LR. The variable component is marked in the name of the LR by the letter X.

Both Extension LR have subordinate Extension LR which further differentiate the structure. LR *wie Musik in X Ohren* can be extended to *wie Musik in X Ohren klingen*. This LR is interesting because it is in fact a combination of two MWE: *Musik in den Ohren* and *in den Ohren klingen [resound in the ears]*. *Musik in X Ohren sein* can be extended to *Das war/ist Musik in X Ohren* and *wie Musik in X Ohren sein*. In fact, the latter LR could also be defined as a child element of *wie Music in X Ohren* as it represents a combination of the two extensions.

Context LR These LR serve as a focus on typical contexts in which the MWE is used. They highlight word forms that commonly appear close to the MWE, but are not directly connected to its structure (as opposed to the elements of the Extension LR). Context LR show the associative frame of the MWE and are thus useful to understand the pragmatics of its usage.

Example: The MWE *Musik in den Ohren* does not have Context LR. However, a typical Context LR would be *Töne ... das menschliche Ohr* which belongs to the MWE *das menschliche Ohr [the human ear]* and subsumes the search patterns `Töne #* das menschliche Ohr` and `das menschliche Ohr #* Töne`. The Context LR highlights a word form, *Töne [sounds]*, which appears significantly frequently in the vicinity of the MWE's Core LR. This is an indicator that the MWE typically refers specifically to the human ability of hearing as opposed to other characteristics of the human ear.

LR tree and LR group The different types of LR can be arranged in a hierarchical structure and may be assigned a narrative comment explaining their specifics. Together, they give a differentiated picture of the MWE in its realizations according to the observed corpus evidence.

Apart from the specialized LR types listed above, there is also the LR group. It serves as a container for collecting all relevant instances of an MWE at once and is always used as the trunk of an LR tree. It captures the overall frequency of the MWE and also preserves instances of realization which are not frequent enough to be highlighted by specialized LR, but may still be of interest for a researcher working with our results.

2.4 The level of MWE

An MWE subsumes an LR tree and is assigned a paraphrase that describes a generalized meaning. The special nature of MWE in our approach becomes evident at this point: An MWE is not a static form, but a complex set of realizations from which a common communicative meaning emerges.

As our approach is based on collocation profiles of target word forms, we established the rule that an MWE must contain at least two concrete and immutable lexical elements – the target word form and one collocation partner. This may seem a somewhat artificial restriction, but it helps greatly in structuring the rich and often overlapping structures that can be detected when analysing language in this manner.

To account for the complex interrelations between MWE, it is possible to define links between them. MWE which have a similar or opposing meaning are connected and their relationships are commented on. Also, structural overlap between MWE is pointed out. Links can be defined between MWE which arise from the same collocation profile (and contain the same target word form), but also between MWE from different profiles.

The example MWE *Musik in den Ohren* is assigned the LR tree shown in Figure 2 and a paraphrase if its general meaning is added: "This MWE is used to express that something is received positively and considered pleasing or beneficial."

Also, links to related MWE are defined. Within the same profile these are for example the MWE *in den Ohren klingen [to resound in the ears]*. As mentioned above, these

two MWE are commonly combined. There is also a link to the MWE *Musik für X Ohren [music for X ears]* which has a quite similar meaning, but is more often used to refer to actual musical preferences. In addition to that, there is a connection to a different profile, that of the singular form *Ohr [ear]* where the MWE *Musik in X Ohr [music in X ear]* can be found, which is nearly identical in meaning, but much less frequent.

2.5 The level of MWE patterns

MWE patterns are abstractions over several MWE. This level accounts for the fact that structures in language can be much more general than the restrictions we imposed on MWE allow for.

Much regard is given in our model to mutable elements in fixed structures, which also appear on the level of surface patterns and LR. On the level of MWE patterns, this concept is brought to a higher degree of abstraction: MWE patterns generalize over MWE which are structurally similar, but different in some aspect of their lexical structure. Therefore these patterns always contain underspecified components.

Two types of MWE patterns can be distinguished. In the first case, the realizations of the underspecified components can be regarded as synonyms. The MWE pattern structurally generalizes over MWE which carry essentially the same meaning.

In the second case, the realizations of the underspecified components are a set of dissimilar lexical items. Each MWE subsumed by the MWE pattern has a different communicative function in its own right, but they share a common core meaning which can be assigned to the more generalized structure.

MWE patterns are especially interesting from the point of view of construction research as they illustrate the transition from concrete lexical items to abstract structures.

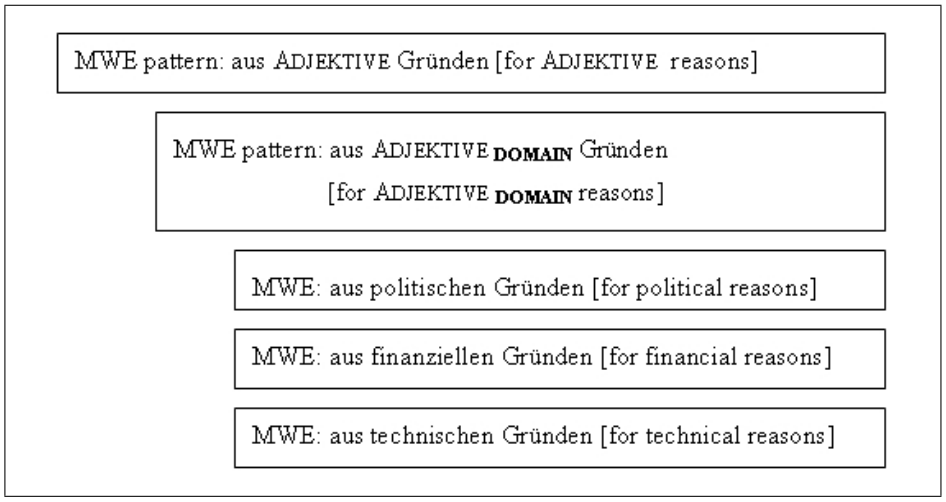


Fig. 3. Excerpt from the hierarchy of the MWE pattern *aus ADJECTIVE Gründen*

An example for an MWE pattern of the second type is the structure *aus ADJECTIVE$_{DOMAIN}$ Gründen [for ADJECTIVE$_{DOMAIN}$ reasons]*. This MWE pattern subsumes several MWE where the underspecified component is realized by a specific lexical item and which each have a different meaning. However, a communicative function which is shared by all its child MWE can be attributed to the MWE pattern: "Using this pattern makes the actions that are explained seem official and at the same time allows the speaker to be vague about the reasons for these actions by using the less specific plural form (*Gründen [reasons]*) which is mandatory for its structure."

Of course, *aus ADJECTIVE$_{DOMAIN}$ Gründen* is itself a specialization of the even more general MWE Pattern *aus ADJECTIVE Gründen*. The meaning assigned to this most abstract MWE pattern is justification in a general sense. Figure 3 shows part of the hierarchical structure of MWE patterns and MWE.

3 Presentation and prospects

In our 'Wortverbindungsfeldern' (MWE fields) [Steyer/Brunner 2008] we have created a hypertext view which reflects the steps of interpretation detailed in this paper and gives access to the corpus data the analysis is based on. In this way, it is possible to completely retrace our interpretative steps and decisions in generalization. The presentation is enriched by comments on the meaning and specifics of MWE patterns, MWE and LR.

At the moment, the focus of our research is on the development of the model of analysis. We are planning to extend the network of interrelations especially between MWE that originate from different collocation profiles and expand the level of MWE patterns. Also, annotations for MWE should be introduced, which would allow dynamic grouping and different views.

Another research question is how to adapt the in-depth method for a larger scale description of MWE and to develop modes of presentation suitable for different user groups – for example learners of German versus researchers interested in constructions.

References

[Belica 1995] Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode. Mannheim. Internet: http://www.ids-mannheim.de/kl/projekte/methoden/ur.html[1]

[Brunner/Steyer 2007] Brunner, Annelen / Steyer, Kathrin (2007): Corpus-driven study of multi-word expressions based on collocations from a very large corpus. In: Proceedings of the 4th Corpus Linguistics Conference, Birmingham. Internet: http://corpus.bham.ac.uk/corplingproceedings07/paper/182_Paper.pdf[1]

[CPA] Corpus Pattern analysis. Internet: http://nlp.fi.muni.cz/projekty/cpa/[1]

[Hanks 2008] Hanks, Patrick (2008): Lexical Patterns. From Hornby to Hunston and Beyond. In: Bernal, Elisenda / DeCesaris, Janet (eds.): Proceedings of the XIII Euralex International Congress, Barcelona, 15-19 July 2008. Barcelona, p. 89-129.

[1] Web sites were checked on 29 July 2009.

[Feilke 2004] Feilke, Helmuth (2004): Kontext – Zeichen – Kompetenz. Wortverbindungen unter sprachtheoretischem Aspekt. In: Steyer, Kathrin (ed.): Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York.

[KLa] Projektseite Ausbau und Pflege der Korpora geschriebener Gegenwartssprache: Das Deutsche Referenzkorpus – DeReKo. Internet: http://www.ids-mannheim.de/kl/projekte/korpora/[1]

[KLb] Projektseite Methoden der Korpusanalyse- und erschließung. Internet: http://www.ids-mannheim.de/kl/projekte/methoden/[1]

[Sinclair 1991] Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.

[Steyer 2000] Steyer, Kathrin (2000): Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. In: Deutsche Sprache 28, 2. Berlin, p. 101–125.

[Steyer 2004] Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (ed.): Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York, p. 87–116.

[Steyer/Brunner 2008] Steyer, Kathrin / Brunner, Annelen (2008): Wortverbindungsfelder. Abrufbar unter "Wortverbindungen online". Internet: http://www.ids-mannheim.de/ll/uwv/wvonline/wvfelder[1]

[Steyer/Brunner 2009] Steyer, Kathrin / Brunner, Annelen (2009): Das UWV-Analysemodell. Eine korpusgesteuerte Methode zur linguistischen Systematisierung von Wortverbindungen. (= OPAL – Online publizierte Arbeiten zur Linguistik 1/2009). Mannheim. Internet: http://www.ids-mannheim.de/pub/laufend/opal/privat/opal09-1.html[1]

[Steyer/Lauer 2007] Steyer, Kathrin / Lauer, Meike (2007): "Corpus-Driven": Linguistische Interpretation von Kookkurrenzbeziehungen. In: Kämper, Heidrun / Eichinger, Ludwig M. (eds.): Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache. (= Studien zur deutschen Sprache 40). Tübingen, p. 493–509.

[Tognini-Bonelli 2001] Tognini-Bonelli, Elena (2001): Corpus Linguistics at Work. (= Studies in Corpus Linguistics 6). Amsterdam/Philadelphia.

Text-Oriented Thesaurus Retrieval System for Linguistics

Natalia P. Darchuk and Viktor M. Sorokin

Kyiv Taras Shevchenko University, Ukraine

Abstract. The aim of the project is 1) a compilation of the electronic Dictionary of linguistic terms using a new formalized thesaurus compilation methodology that meets the standards of the terminography and its presentation in Internet; 2) a verification of the theoretical thesaurus model owing to the developed computer technologies by using it for the analysis of texts taken from various spheres of linguistics.

The importance of the project is that IRS in the multimedia space first of all provides linguists with a modern standardized dictionary of linguistic terms; secondly, the result of the project is a methodology of thesaurus compilation and also computer tools for realization of the methodology; thirdly, IRS is compatible with intellectual systems of text information processing where it can be implemented as a knowledge base and a tool for text meaning recognition.

Logic-notional modeling of terminology systems in different sciences and branches of knowledge is one of the topical inter-branch problems nowadays as the models of terminology systems are required for compiling terminological dictionaries, information thesauruses, classifiers, rubricators, creating automated retrieval systems, databases, artificial intelligent systems. Development of the information-retrieval thesaurus (IRT) that, on the one hand, is a mean for formalized representation of terminology, since it represents semantic relations between terms rather strictly, and, on the other hand, is an important tool for constant improving of knowledge systems of specific sciences, is considered to be a particular case of knowledge modeling. The project that is developed in the laboratory of the computational linguistics of the Philology Institute in the Kiev National Shevchenko University is dedicated to this problem.

The aim of the project is

- a compilation of the electronic **Dictionary of linguistic terms** using a new formalized **thesaurus compilation methodology** that meets the standards of the terminography and its presentation in Internet;
- a verification of the theoretical thesaurus model owing to the developed computer technologies by using it for the analysis of texts taken from various spheres of linguistics.

The project was divided into two phases. At the **first stage**, we developed information retrieval system represented as a lexicographic and encyclopedic electronic database of the Ukrainian linguistic terms that consists of four dictionaries: **alphabetic, translation, explanatory and thesaurus.** In the **alphabetic dictionary** every term (3400 terms) or terminological word combination has its **English equivalent** (it is planned to add Russian, German, French and Italian) and definition (Figure 1). **Thesaurus** is a list of logic-semantic relations between linguistic terms (the list was taken from the work [3] but we supplemented and modified it). So, the developed IRS includes not only separate terms represented as an alphabetically arranged list together with their interpretations but also the models that show the relations between terms (Figure 1).

On the basis of the modern linguistics achievements interpretations of the terminological units are taken from the reliable sources (about 30 of them) such as terminological dictionaries, grammars, monographs and given in the compact and available form. The dictionary includes general linguistic terms – mainly nouns or noun phrases – from all spheres of grammar, lexicology, applied and computer linguistics.

Dictionary units are widely used general linguistic terms from morphemic, word formation, paradigmatic, syntax, lexicology, semantics, terms from particular applied spheres known in the Ukrainian, Russian and foreign linguistics, terms from computer linguistics connected with the automation of linguistic processes. Selection of the dictionary units to the database was based on the heuristic principles (knowledge of the thesaurus compilers and linguistists-experts).

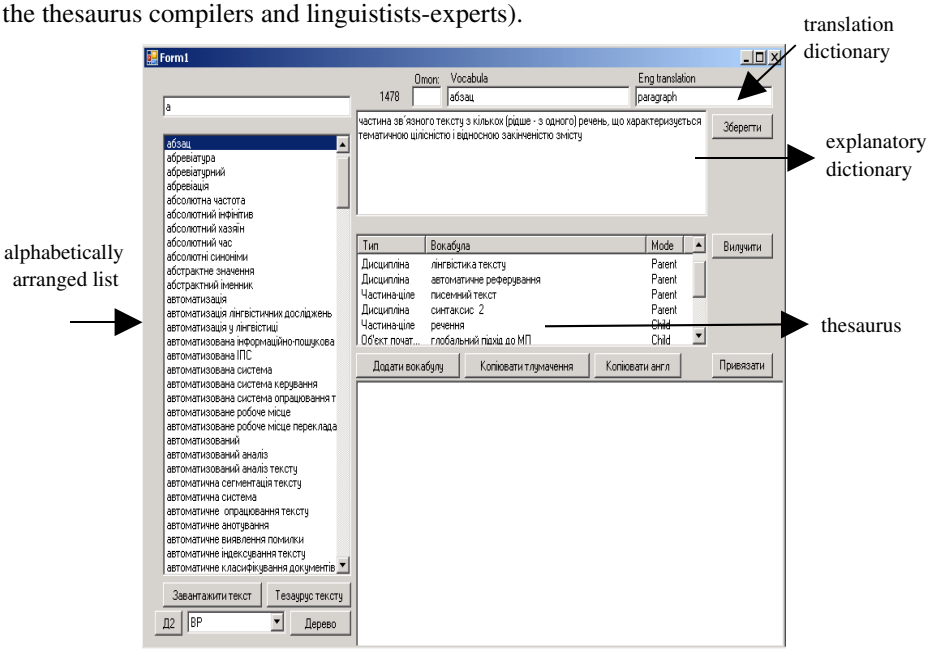


Fig. 1. Fragment of the electronic dictionary of linguistic terms

The process of the thesaurus compilation is supposed to reveal all types of relations between notions expressed by terms. The main relations are hyponymy (genus-species), collateral subordination at one level – partiation (whole - part), synonymy, correlation, association, object localization, its destination, function, means of function expression, relations etc. Content of relations is widened so, that it is possible to cover maximal number of terms connected with the analyzed term as a dictionary unit. As an interpretation was sometimes not enough to get all essential for the terms relations, we also took into consideration encyclopedic dictionaries, scientific works on the particular problem, our own knowledge and knowledge of linguists-experts. Dictionary entry is structured as a form that was filled for every term. The form includes a standard list of relations that are notional for a dictionary unit. The name of a relation is a binary predicate R (A, B) that links the main word of the entry (A) with the term (B) introduced by this predicate [3, 22].

Thesaurus consists of **3394** terms that are covered by a semantic net with **9265** semantic relations (Table 1).

Type of semantic relations	**Term A (example)**	**Term B (example)**	Frequency of semantic relations' realization in the thesaurus
Genus – Species (***A is generic to B***)	*частини мови*	*несамостійні частини мови*	1843
Discipline (***A розглядається в дисципліні B***) (A is regarded in discipline of B)	абзац	синтаксис	1157
Synonyms (***A is synonymous to B***)	автоматичний переклад	машинний переклад	1016
Whole - Part (***B consists of A)***	абзац	писемний текст	934
Correlate (***A is opposite to B***)	агент	пацієнс	822
Watch…(***about A watch B***)	акузатив	знахідний відмінок	529
Association (***A is associated with B***)	субморф	морф	312
Operation/Procedure (***B is an operation/procedure for A***)	абревіатура	абревіація	263
Initial object (***B is performed over A***)	речення	автоматична сегментація тексту	232
Tool/Method ***(A using B***)	автоматичне анотування	мережеве моделювання	221
Parameter (***A is characterized by B***)	аломорф	додаткова дистрибуція	205

Parameter carrier (***B is a carrier of parameter A***)	абревіація	іменник	210
Final object (***A is performed over B***)	автоматичний синтез тексту	текст	202
Is related to…(concerns adjective terms: ***A is related to B***)	акузативний	акузатив	189
Way of expression (***A is expressed by B***)	особове дієслово	особова форма дієслова	184
Language level (***A is regarded at the level specified by B***)	актантна структура	синтаксичний рівень	176
Main function (***A expresses B***)	*вигук*	експресивність	157
Relation (***B restricts A***)	*головний член*	*безпосередня синтаксична залежність*	130
Way of the object representation (***A is presented through B***)	словотвірна структура слова	деривaційна історія слова	109
Aspect (***A is regarded in the aspect B***)	графіка	*писемний текст*	99
Linguistic object (**A *is presented as B***)	*зв'язний текст*	*семантична мережа*	96
Implication (***if A, then B***)	*знахідний відмінок*	*перехідне дієслово*	81
Unit of the level (***B is a unit of A***)	*морфологічний рівень*	словоформа	40
Class (***A is a member of class B***)	аломорф	морф	36
Object of a science (***A is an object of B***)	*природна мова*	*інженерна лінгвістика*	22

Table 1. Quantitative characteristics of semantic relations

The main paradigmatic semantic relations (genus-species, synonymy, whole-part, correlation), as it can be seen from the table, cover considerable amount of terms (almost 70% of all paradigmatic relations). From the point of view of theoretical semantics the more semantic information a dictionary includes, the better, because the large number of relations in the thesaurus gives a user a greater opportunity to express his information needs in the request. As the dictionary entry is a synthesis of linguistic, translation, explanatory and encyclopedic information, we, in connection with information approach intended for a user request in the IRS, developed an interactive system (the example is given below):

Request: «what term is a generic for a term "***абстрактний іменник***" (*abstract noun*) ?»

Answer: «A noun "***іменник***" (*noun*) is a generic».

Request: «What terms and relations is a term "***абстрактний іменник***" (*abstract noun*) related to?»

Answer: «"***Граматичний рід***" (*grammatical gender*) – relation: ***parameter carrier;*** "***Абстрактне значення***" (*abstract meaning*) – relation: ***parameter carrier, implication***»

The answers to the request are taken from the thesaurus graph that looks like a semantic net – hierarchically arranged data structure of **nodes** (terms) and **ridges** that represent different types of thesaurus relations – and is given from the thesaurus as a text (Figure 2).

Relations "node (n) – servant (s)" where a dictionary unit is a "node" that subordinates are colored black on the screen, and relations "servant - node" where a dictionary unit is vice versa subordinated are colored blue.

There are terms in the thesaurus that are covered by a ramified network of semantic relations: *sentence* – 121 (n = 24, s = 97); *word* – 62 (n = 30, s = 32) etc.

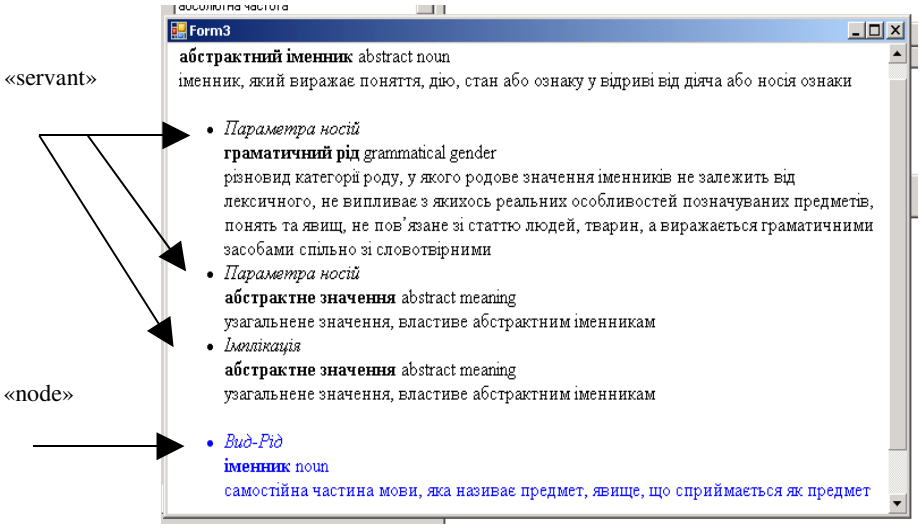


Fig. 2. Thesaurus graph

Since every relation has its own mark and number in the electronic format, it facilitates getting data grouped the way a user wants. Represented thesaurus model can be regarded as a semantic and information description of linguistic terminology that is modified as

- alphabetically arranged dictionary with definitions;
- sectional terminological dictionary of such disciplines as morphology, cognitive science, computer linguistics etc;
- dictionary of synonyms;
- dictionary of genus-species relations;
- automatic information reference book etc

Network data representation has not only applied necessity but also enables deeper understanding of logic relations in this science, more precise modeling of the analyzed terminological system.

Compiled thesaurus is **a static model** of logic-notional relations between terms of linguistic metalanguage. This model can be examined from the point of view of representing **dynamic aspects** of the scientific knowledge structure, meaning as a verification of theoretical thesaurus model using it for the corpus analysis of texts from different spheres of linguistics. The importance of such research is explained by the fact that knowledge has a textual representation and becomes known from texts. Text is a different form of the existence of scientific knowledge. Here knowledge functions as a text semantics, that is knowledge gets gradually a status of scientific information [1, 50]. There is no doubt that the most complete knowledge about an object, its features and properties are given in the genre of scientific monographs such as textbooks, dissertation, because inter-branch relations are examined in details in the scientific monographs. However, other types of texts such as report papers, articles, discussions etc, that represent scientific knowledge in a different way, cannot be disregarded.

Any encyclopedic model of the scientific knowledge is a derivative from a set of real texts and a representation of these texts at the level of the semantic model. Semantic model of the language of science is an invariant of separate partial texts' micromodels relatively to which texts are its variants. Relatively to the set of terms of the particular science, a logic-notional system of the knowledge field is a model of the plane of content of the knowledge field. The logic-notional structure of the text represents the main elements of a semantic paradigmatics of texts. Modeling of scientific texts' semantics supposes revealing all types of semantic relations among all words – notions' names, their features and properties etc.

Two approaches that give different theoretical and practical results are applied while investigating a term in the text: from the text to the term (that is terminological text analysis) and from the term to the text (textual term analysis). Terms are formed in the text but registered in the terminological system (dictionary). And if in the functioning sphere the term exists as paradigmatic and syntagmatic variants then in the terminological system it exists only as paradigmatic variants [2, 152].

On developing methods according to which created thesaurus model of linguistic terms represented as a hierarchical classification scheme-graph is imposed upon the vocabulary of a scientific text, a combination of these approaches is supposed to be promising. As a result we also get a hierarchical classification graph of the specific analyzed text with absolute frequency of a term usage in the specific text. This analysis is provided by the second approach – from the terminological dictionary to the text. Next methodical step is development of an auxiliary list of words on the basis of the text together with corresponding frequencies without taking into consideration those terms that were included to the thesaurus graph. Revising the words of the list it is possible to form a list of words that can be considered to have all the potential to

gain a status of terms. This is provided by the approach from the text to the term, and the frequency of usage and contexts enable accumulating scattered terminological information to solve different terminological problems, in particular to analyze the reasonability of adding words-pretenders to the compiled dictionary of terms.

This explains developing of **the second stage** of this project – creating the dynamic logic-notional model of each particular text from the corpus with texts of linguistic research area.

The methodology consists of the following main steps: 1) lemmatization and arrangement of lemmas according to their parts of speech; 2) calculation for each lemma (noun and adjective) of its absolute usage frequency; 3) creation of the thesaurus graph of terms for a specific text with their absolute usage frequencies in the text applying the thesaurus graph of the term system; 4) resolving homonymy of terms' meanings (e.g. **grammar 1** *(a systematic description of the grammatical facts of a language, that is a system of morphological units, categories and forms, syntactic units and categories, derivational units and means of word formation)*; **grammar 2** *(the branch of linguistics that deals with grammatical structure of a language)* and **grammar 3** *(formal grammar that is a part of the automatic text processing system)*) with the help of a context; 5) compilation of an auxiliary word list with absolute frequencies that was not included to the thesaurus; 6) search for the words-pretenders on the role of the terms and their registration together with illustrative context in the additional terminological dictionary. The work of steps 1-3 and 5 in automatic mode and 4 and 6 in automated mode is provided by the algorithm and program packages. The performance of the methodology was checked on the corpus of linguistic scientific articles, because the article as one of the forms of scientific knowledge existence discusses one of the aspects, shows one of the sides of the scientific notion or problem providing a wide range of scientific knowledge, its represented terminological semantics (texts of 72 articles from the collection "Українське мовознавство"; text length is 82418 word usages).

As the result of automatic applying of the thesaurus represented as an oriented graph to the general word list of the corpus with the linguistic scientific articles we got another word list also represented as an oriented graph. Considering previously compiled thesaurus as an invariant model of the linguistic knowledge, relatively to the analyzed texts thesaurus is a plane of content model: semantic net represents hierarchically organized data structure consisting of the nodes (terms) – the notions' names – and ridges that show the relations between these notions (Figure 2).

Word list of 821 terms covers 16,12% of text (13287 terms of 82418 word usages in the text cumulatively). At the same time about 24% of linguistic terms included to the thesaurus were found in the corpus (821 terms of 3340 terms in the thesaurus).

The next step was checking whether the automatic auxiliary word list contains terms or terminological word phrases that were not found in the general linguistic thesaurus because of the number of reasons: firstly, there is can be author's terms included to the text because of the necessity to show the author's position about some

theoretical problem; secondly, dictionary can never be complete as the process of science-technological terminology development can never be stopped. Text is a source where a term is formally and semantically formed. Only in the text it is possible to retrace “the life” of the term solving the question about reasonability of including it into the terminological dictionary.

Described methodology enables finding and further examining of the term functioning in the automated mode. Terms in texts are formed together with ideas, notions and are followed sometimes with such phrases as “by this term something is meant…” etc. This can be a reason for compiling a list of words-pretenders on the role of a term together with their contexts enough for solving the problem of reasonability, or, in case of giving different from generally accepted interpretation, a the reason for saving them together with the interpretation in the specially developed terminological data bank (TDB)

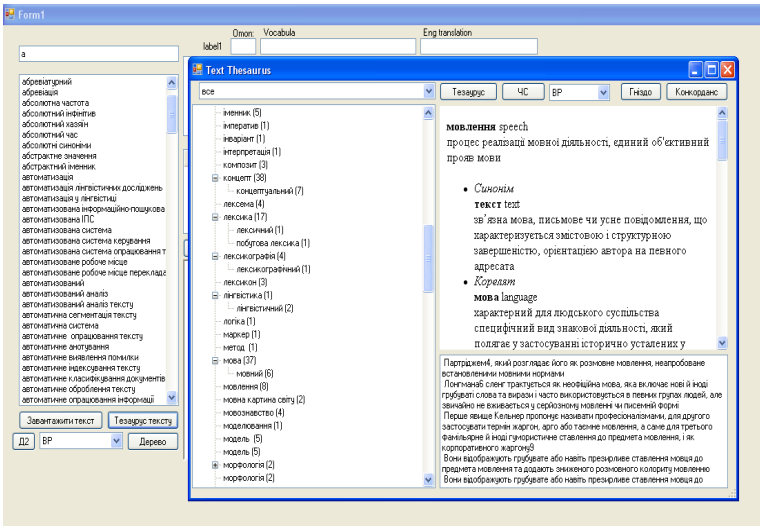


Fig. 3. Thesaurus of terms from the collection of articles «Українське мовознавство»

The auxiliary word list contains terms from the following branches of linguistics:

Stylistics: *фоностилема, стиледиференторм, іностильове включення, ритмічні синтагми, ритмізація синтагми, горизонтальна / ланцюжкова ритмізація текстів, вертикальна ритмізація тексту* (В.Берковець).

Cognitive science: *домовна картина світу, просторова картина світу, національно-мовна картина світу, міжнародно-правова картина світу, асоціативний ґештальт* (І. Скорик); *ядро ґештальту, зони ґештальту* (Д.Терехова); *парадигма мислення, когнітивна модель, когнітивна діалектологія* (Н.Руснак), *інгерентний аспект конотації, адгерентний аспект конотації* (І.Калита).

Lexical science *категорійна система лексикону школяра* (В.Чумак, А.Акуленко); *колоквіалізм* (І.Ковалинська); *народні компаративеми, еліптування* (О.Дуденко); *інтерпретаційна модель оксюморонів* (В.Мерінов); *загальний сленг, спеціальний сленг* (В.Вітренко); *уніноміальні назви вищих таксонів* (С.Руденко).

Semantics: *субкоди семантичної субстанції* (Л.Скиба), *психосемантика* (Т.Печончик).

Syntax: *хронотип, компаративема* (В.Коломийцева); *граматика поетичних інновацій, експресивний синтаксис* (Н.Гуйванюк).

Morphology: *міноративні (форми, суфікси), міноративність* (Н.Руда); *аглютинант* (С.Климович).

Individual author's terms can also be found sometimes. But formed once individual terms can become a social phenomenon if the theory, where this term appears, becomes generally accepted being practically confirmed (e.g. the term "performative verbs" by John Ostin is used in the papers of the Ukrainian grammarian I.R. Vuhovanec and in the article written by I. Shevchyk.)

The research showed that terms that were not included into the thesaurus are either rarely used in linguistics (*ґештальт, хронотип, таксон, субфрейм, суперфрейм*), or author's ones (*міжнародно-правова картина світу, домовна картина світу, колоквіалізм, народні коппаративеми, психосемантика* etc), and in any case they were included with definitions provided by the authors of the scientific articles.

Such ways of word formation can be retraced among one word innovations:

- stem-composition (*фоностилема, стиледиференторр, психосемантика, хронотип*): an object is designated according to the set of features;
- usage of Greek-Latin patterns (*компаратив-**ем**-а,колоквіал-**ізм**, фоностил-**ем**-а, аглютин-**ант**, міфолог-**ізм**);*
- abbreviation (***ККС** – концептуальна картина світу, **ПКС** – просторова картина світу, **РС** – ритмічні структури, **ЕКС** – емоційно-смисловий концепт*).

There are more two-component nominative word combinations that consist of a designated word (usually a noun) and a determinant word (adjective, another noun). Multicomponent word combinations tend to transform into the abbreviation (*МКС – мовна картина світу*). In such combinations a determined component is usually a name of specie that is differentiated by adding a determinant:

1. ***картина світу (worldview);***
2. *міжнародно-правова **картина світу** (international **worldview**);*
3. *концептуальна **картина світу** (conceptual **worldview**);*
4. *національно-мовна **картина світу** (national language **worldview**);*
5. *двомовна **картина світу** (bilingual **worldview**);*
6. *просторова **картина світу** (spatial **worldview**).*

The main term (genus: *картина світу (worldview)*) and derivative from terminological element, that determines a feature of the specific notion, forms the whole terminological paradigms complicating the basic notions that are lately being used in the cognitive science papers.

Authors used the following methods while term forming:

1. term formation on the basis of the international terminological elements – morphological components;
2. loanword of terms – usage of transliterated foreign terms. That means that the same laws are applied both for term formation and for the formation of any other lexical units of the Ukrainian language. The typical approach is term combination.

Let's consider the structure of the terminological thesaurus compiled on the basis of the linguistic articles and some of the results. It should be mentioned that percentage of specific and non-specific vocabulary in the word list is approximately 21,3% of the non-terms – generally used words and 78,7% of the terms. Such high percentage of terms evidences that the chosen texts are using and fixating the terms. Moreover, architectonics of the text should also be taken into consideration. Architectonics means a typical structure of the article as a textual genre where the subject, preconditions, reasons of the research, novelty, author's position, sources of the research, experiment methodology, experiment description, results of its discussion, examples, conclusions, results are highlighted. That is why the modeling of semantics was oriented towards previous logic-notional, conceptual content analysis of the scientific papers (not only dictionaries but also monographs, textbooks, articles were used during thesaurus compilation before analysis). The thesaurus represented as an alphabetical-frequency tree where terms are the nodes and the points head subtree that consists of the nodes of the higher levels of the hierarchy (total absolute frequency either of the whole subtree or the frequency of the particular term in the branch is given in the brackets).

For example,

Аналіз (95) (Analysis)
Аналіз тексту (1) (Text analysis)
КА (СА) (6) (абревіатура: контекстний аналіз) (abbreviation: context analysis)
Концептуальний аналіз (1) (Conceptual analysis)
Логічний аналіз (1) (Logic analysis)
Семантичний аналіз (1) (Semantic analysis)

The term *analysis* heads the subtree and its absolute frequency (95) includes the frequencies of other nodes at the lower levels of the hierarchy. So, the frequency of this term only is 85.

The frequency terminological dictionary (its upper part) served as the subject model. The most frequent terms represent the subject dominant of texts. They are *мова (language)* (624), *слово* (*word*) (360) *форма (form)* (345), *значення*

(meaning) (284), *семантика* (284) *текст* (249) (*text semantics*), *мовлення (speech)* (213), *структура (structure)* (205), *іменник (noun)* (195), *лексика (vocabulary)* (183), *ознака (feature)* (172), *прикметник (adjective)* (143), *предикат (predicate)* (134), *речення (sentence)* (133), *система (system)* (168), *суфікс (suffix)* (153), *функція (function)* (163) etc.

The terminological system of the investigated texts includes several groups of terms. First of all, they are basic terms of methodological sciences, for example, of **philosophy** (*аналіз (analysis)* – 95, *відношення (relations)* – 80, *система (system)* – 168, *форма (form)* – 345), **logic** (*концепт (concept)* – 154, *суб'єкт (subject)* – 57, *об'єкт (object)* – 96, *адресат (addressee)* – 55, *фрейм (frame)* – 23), **computer science** (*інформація (information)* – 69, *функція (function)* – 163, *конструкція (structure)* – 92, *модель (model)* – 78). The terminological system includes the main terms that designate the notions of the **linguistic branch** (*іменник (noun), прикметник (adjective), дієслово (verb)* (71), *речення (sentence), лексема (lexeme)* (92), *слово (word)* (360) etc.), and its derivatives (e.g. *основа (basis)* (240): *граматична основа (grammatical stem)* (1), *лексична основа (lexical stem)* (1), *основа дієслова (verb stem)* (1), *основа іменника (noun stem)* (1), *основа слова (stem)* (2), *основа словоформи (wordform basis)* (1), *складна основа (complex basis)* (2), *непохідна основа (underived basis)* (1), *похідна основа (derived basis)* (19), *основоскладання (stem composition)* (3)) that form lexical terminological paradigm represented by a word combination.

So-called general scientific terms don't differ from the highly specialized terms of the methodological sciences by their form, but semantics of general scientific terms differentiates getting additional semes.

Модель (Pattern) (general scientific term) (78)
Мовні моделі (Language patterns) (2)
Моделі мовленнєвої діяльності (Speech activity patterns) (1)
Модель речення (Sentence pattern) (1)
Синтаксична модель (Syntactic pattern) (1)
Словотвірна модель (Word-building pattern) (1)

In the terminological vocabulary of the linguistic articles the number of one word terms (524) is twice as large as the number of two word terms (269) and three word terms (22). That can be explained by the fact that the terminological system of linguistics is an old one as well as linguistics itself. However, the expansion of the system due to the new word combinations mainly of the *adjective + noun* type occurs as the result of the science development and mutual integration of different subjects. This tendency occurs when there is an absence of clarity while designating some notions with one word. In this case additional components are being added to it. If designation has redundant clarity or accuracy then it is being cut forming abbreviation (e.g. *translation – machine translation - MT*).

Thesaurus representation of the terminological system as a semantic net is also a part of the science language description. IRT is a lexical structure pattern as the

thesaurus reflects the paradigmatic structure of vocabulary in an explicit form and enables systematizing all main vocabulary of the investigated material (corpus of the linguistic articles in this case) from the paradigmatic point of view.

The thesaurus illustrates semantic continuity: dictionary cannot contain semantically isolated terms. Every term is related to the meanings of the other by the relations of genus-species, partiation, synonymy and correlation. It is possible to move from the semantic net to the subject one on the following bases: the more terms (with their quantitative indexes) in the text are connected in the thesaurus tree by the semantic relations, the greater subject significance they have in the text. Subjects are the names of the subject groups in the thesaurus tree. On the other hand, the subject net is a hierarchy of the main and additional subjects. The following methodological statements are the initial: 1) thematically significant terminological groups are being developed on the basis of the thesaurus tree; 2) the hierarchy of subjects of the text is determined depending on the number of terms included into subject groups.

Addition of a new scientific text in the linguistic corpus causes updating of the IRT model. This model enables: 1) automatic selection of the terminological lexemes on the basis of a coincidence with a general thesaurus; 2) automated control provided by a specialist: identifying terms that are not included into the thesaurus; giving them definitions, translations, establishing thesaurus connections.

The process of lexico-subject model development consists of the three stages:

1. lemmatization of the text word forms, compilation of the frequency word list of the text and its arrangement according to the parts of speech;
2. calculation of the absolute frequency that shows the number of terms met in the branch and the general frequency of the term in the texts;
3. homonymy resolution of the terminological lexemes and enumeration of the frequency characteristics.

Significance of the corpus research of the terms' functioning is beyond dispute, as scientific theories, concepts are outlined and methods are described in the texts. And terms are then included into dictionaries (or standards, classifiers). So, the main principle in the science of terminology is the initial sphere of term functioning in so-called term forming and term using texts.

The following methods were used during terminological text analysis and textual term analysis:

- logic method of classification and construction of the terminological system establishing relations between terms;
- logic, linguistic, terminological methods applied to identify the facts of combined usage of terms together with non-terms and general scientific terms, analyze the combinations of terms with terms in the text;
- quantitative methods applied to identify the importance of the term.

Compiled text-oriented IRT with linguistic instrument and software can be applied while developing terminological data banks (TDB) and terminological knowledge banks (TKB), since a large amount of terminological information has been worked out (3400 units), tools to automate the processes of selecting, keeping and searching this information: dictionary of linguistic terms is kept in the computer memory and can be developed and supplemented by new words. Advantages of such dictionary of a traditional paper one lie in the fact that electronic IRT enables entering through any characteristic of a term. On the other hand, the term is not restricted by its dictionary entry like in the paper dictionary – logical and associative relations are established between terms in the TDB: relations of genus-species, partiation, synonymy, correlation etc. Semantic nets, that explicitly represent the structure of the multilevel terms' classifications within branch linguistic terminological system, are developed with the help of the programs.

The importance of the project is that IRS in the multimedia space first of all provides linguists with a modern standardized dictionary of linguistic terms; secondly, the result of the project is a methodology of thesaurus compilation and also computer tools for the methodology realization; thirdly, IRS is compatible with intellectual systems of text information processing where it can be implemented as a knowledge base and a tool for text meaning recognition (automatic abstracting and annotating of the scientific linguistic texts).

References

[1] Герд А. С. Прикладная лингвистика. – Издательство С.-Петербургского университета. – 2005. – 266 с.

[2] Лейчик В. М. Терминоведение: Предмет, методы, структура. – М.: КомКнига, - 2006. – 256 с.

[3] Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике. – М., 1978.

From Electronic Corpora to Online Dictionaries (on the Example of Bulgarian Language Resources)

Ludmila Dimitrova

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

Abstract. The paper briefly describes Bulgarian digital language resources, among them corpora, lexical databases, lexicons, and electronic dictionaries, which were developed in the Mathematical Linguistics Department at the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences (IMI-BAS) in the framework of some international projects. The first Bulgarian electronic corpora and language-specific resources were developed in the EC language technology project MULTEXT-East Multilingual Text Tools and Corpora for Central and Eastern European Languages. The first lexical database for Bulgarian was developed in the EC project CONCEDE Consortium for Central European Dictionary Encoding. These recourses were developed in TEI-format, and thus they are compatible with other TEI-conformant resources. The first Bulgarian-Polish electronic corpora and dictionaries are currently developed in the frame of bilateral collaboration between IMI-BAS and ISS-PAS.

1 Introduction

The Department of Mathematical Linguistics at the IMI-BAS has successfully participated in the EC language technology projects MULTEXT-East and CONCEDE. The MULTEXT-East project (MTE for short: [2]), as a continuation of the project MULTEXT *Multilingual Text Tools and Corpora* [10], aims at testing and adaptation of language standards and corpus tools, developed through the MULTEXT, the development of language-specific resources for six new languages, and the extension of the annotated multilingual MULTEXT corpus. MTE developed digital language resources for six Central and East European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, as well as for English. These resources contains, for the each of the six CEE languages, morphosyntactic specifications, lexica, and corpora. The corpora consist of three parts: parallel, comparable and speech-corpus. Developed in the frame of the MTE project, Bulgarian language digital resources include morphosyntactic specifications, lexicons, and corpora, incl. a parallel corpus, based on George Orwell's novel 1984, and a comparable corpus [6].

2 Bulgarian language-specific resources: morphosyntactic specifications

The MTE morphosyntactic specifications have been developed on the basis of the MULTEXT specifications for Western European languages and in accordance with the EAGLES guidelines [11]. The morphosyntactic specifications have been used in the encoding of the word-form lexica of the project. They contain the list of defined

categories – parts of speech (POS), each POS encoded by a letter: noun – N, verb – V, adjective – A, pronoun – P, determiner – D, article – T, adverb – R, adposition – S, conjunction – C, numeral – M, interjection – I, residual – X, abbreviation – Y, particle – Q. A table of attribute-values is defined for each category in order to reflect the characteristic features of each language. The specific features of each language are marked up additionally by **l.s.** The characters following the POS-encoding give the values of the position-determined attributes. The specifications define, for each part of speech, its appropriate attributes and their values, encoded by one symbol code. It should be noted that if a certain attribute is not appropriate (1) for a language, (2) for the particular combination of features, or (3) for the word, this is marked by a hyphen in the attribute's position.

The MTE use the MULTEXT format of lexical description – morphosyntactic description (MSD), which consists of linear strings of characters, representing the morphosyntactic information for each word-form. The string is constructed in the following way:

- the positions of a string of characters are numbered 0, 1, 2, etc.
- he agreed character at position 0 encodes the corresponding part of speech: N for noun, V for verb, etc. ;
- each character at position 1, 2, n, encodes the value of one attribute (for nouns the attributes are: type, person, gender, number, etc.);

For example, the MSD **Ncfs-** means POS: noun, Type: common, Gender: feminine, Number: singular, nocase.

The proposed formalism for the MSD is not arbitrary (a MSD contains the full description of a lexical item), but has a clear and concrete aim – to be used for specific applications, incl. corpus annotation. On the basis of these standard MSDs the set of corpus tags were determined. A mapping from the morpho-syntactic information, contained in the lexical description, to a set of corpus tags is also provided, according to the MULTEXT tagging model. The list of MSDs for Bulgarian contains 326 elements.

For example, the MSD of the word **картата** */the map/* is **Ncfs-y** that means POS: noun, Type: common, Gender: feminine, Number: singular, nocase, Definiteness: yes.

Some of MSDs for Bulgarian are not strictly adequate to the particular morhosyntactic properties of the respective parts-of-speech – Tense, Number, Gender, Voice, Definiteness – especially in the system of impersonal verbal forms (participles) [12]. In particular, the *present active participial* cannot possess the Tense attribute because it expresses the property/attribute independently and regardless of the tense of the main verb in the sentence, whereas the Voice attribute is also implicit from the context. New MSDs for Bulgarian participles have been proposed, bringing the morphosyntactic description in line with the grammatical characteristics of the Bulgarian, [7]. An update of the MSDs will make them more useful for annotation of corpora and automatic disambiguation of Bulgarian texts.

The Bulgarian language-specific resources also include a set of segmentation and morphological rules and data, which are necessary for use with the various annotation MULTEXT tools. Segmentation rules describe the form of sentence boundaries, quotations, numbers, punctuation, capitalization, etc. Morphological rules, needed by the morphological tools, provide exhaustive treatment of inflection and minimal derivation. The so called special tokens, required by the segmenter, includes lists of special tokens (frequent abbreviations and names, titles, patterns for proper names, etc.) with their types. Since some subtools, for example, in the segmenter require certain language-specific information in order to accomplish their tasks, each participating side has developed a set of resource files for their language. For maximum flexibility and to retain language-independence, all such information is provided directly to the subtools via external resource files.

3 Corpora

MTE is building an annotated multilingual corpus, composed of three major parts: **Parallel Corpus, Comparable Corpus**, and a small **Speech Corpus** of spoken texts in each of the six languages, comprising forty short passages of five thematically connected sentences, each spoken by several native speakers, with phonemic and orthographic transcriptions. The multilingual parallel corpus, based on George Orwell's novel "1984" in the English original and the six translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene of the novel, was developed. The parallel corpus is produced as a well-structured, lemmatized, CES-corpus [9]. The corpus contains four parts, corresponding to the different levels of annotation: the original text of the novel, the CesDOC-encoding (SGML mark-up of the text up to the sentence-level), the CesANA-encoding (containing word-level morpho-syntactic mark-up), and the aligned versions in CesAlign-encoding (containing links to the aligned sentences). The texts were automatically annotated for tokenization, sentence boundaries, and part of speech annotation, using the project tools, and validated for sentence boundaries and alignment. The alignment between the English version and a translation in each of the six CEE languages ensures six pair-wise alignments.

The next examples show excerpts of the ***Bulgarian-English aligned texts*** – Bulgarian-English Aligned 1984 Sampler:

1-1 Aligned sentences:

- <Obg.1.1.7.4>Още три сгради, подобни по външен вид и размери, бяха посети из **Лондон**.
- <Oen.1.1.9.2>Scattered about **London** there were just three other buildings of similar appearance and size.
- <Obg.1.1.7.5>И дотолкова се извисяваха над околните здания, че от покрива на жилищен дом **Победа** можеха да се видят и четирите едновременно.
- <Oen.1.1.9.3>So completely did they dwarf the surrounding architecture that from the roof of **Victory Mansions** you could see all four of them simultaneously.

1-2 Aligned sentences:

- <Obg.1.1.7.3>От мястото си **Уинстън** можеше да прочете изписани с елегантни букви върхубялата фасада трите лозунга на партията: "Войната е мир""Свободата е робство""Невежеството е сила" Говореше се, че в **Министерството на истината** има три хиляди стаи над земята и съответните лабиринти отдолу.
- <Oen.1.1.7.3>From where **Winston** stood it was just possible to read, picked out on its white face in elegant lettering, the three slogans of the **Party**: "War is peace""Freedom is slavery""Ignorance is strength."<Oen.1.1.9.1> **Ministry of Truth**, contained, it was said, three thousand rooms above ground level, and corresponding ramifications below.

Bulgarian MTE parallel corpus

The Bulgarian parallel corpus contains the ***Bulgarian translation*** of Orwell's novel "Nineteen Eighty-Four", includes 86020 words (lexical items, excluding punctuation), 101173 tokens (words and punctuations); the ***CesDOC-encoding*** of the Bulgarian text of the novel (SGML mark-up of the text up to the sentence-level) includes 1322 paragraphs, 6682 sentences; the ***CesANA-encoding*** of the Bulgarian text of the novel (containing word-level morpho-syntactic mark-up), and the ***Bulgarian-English aligned texts*** – the aligned versions in ***CesAlign-encoding***, containing links to the aligned sentences (*see* examples bellow). The ***CesANA-encoding*** for Bulgarian in addition includes disambiguated lexical information for the 86020 words of the novel and undisambiguated lexical information for 156002 words. What is more – there are 156002 occurrences of MSDs in the text (Bulgarian MSD are 326) and 242 022 occurrences of base or lemma of tokens (which is the total of 86020 words and 156 002 occurrences of MSD). The number of occurrences of ctags is 257 175. Each word-form is associated with the respective grammatical information and the corresponding lemma which form its standard lexical description.

An example of the ***CesANA-encoding*** of the Bulgarian text of "1984" follows:
word-form "**и**" */and* – conjunction/, */so! oh!* - interjection/

```
<tok type=WORD from='Obg.1.1.1.1\24'>
  <orth> и </orth>
  <disamb><base> и </base><ctag>CC</ctag></disamb>
  <lex><base> и </base><msd>Ccs</msd><ctag>CC</ctag></lex>
  <lex><base> и </base><msd>I-s</msd><ctag>I</ctag></lex>
</tok>
```

Bulgarian-Polish parallel corpus

The first Bulgarian–Polish corpus (currently under development) is a result of the joint collaborative project "Semantics and contrastive linguistics with a focus on a bilingual electronic dictionary" between IMI—Bulgarian AS and ISS—Polish AS, coordinated by L. Dimitrova and V. Koseska. It contains a total of approximately 5 million words and comprises two corpora: parallel and comparable [3]. The first

Bulgarian–Polish parallel corpus contains more than 3 million words mainly works of Bulgarian and Polish authors – short stories, novels, children's literature, science fiction. A small part comprises official documents of the European Commission available through the Internet. The corpus is composed of two parts: original Bulgarian texts with Polish translations or *vice versa* and texts in other languages translated into both Bulgarian and Polish.

The corpus is developed according to the MTE model. Most texts have been annotated at paragraph level. The corpus provides samples for the experimental version of the Bulgarian–Polish digital dictionary.

In the framework of the joint collaborative project „Electronic corpora – contrastive study with focus on design of Bulgarian-Slovak digital language resources" between IMI—BAS and ĽŠIL—Slovak AS, coordinated by L. Dimitrova and R. Garabík a small ***Slovak-Bulgarian parallel corpus*** is currently under development.

A small ***parallel corpus*** with Bulgarian, Polish, Slovak, Slovene (incl. English as a hub language) texts of official documents of the European Commission available through the Internet is also currently collected.

Bulgarian comparable corpora

Bulgarian MTE comparable corpus: For each of the six MTE CEE languages, a comparable corpus was developed. It included two subsets of at least 100,000 words each, consisting of

- fiction, comprising a single novel or excerpts from several novels;
- newspapers.

The data was comparable across the six languages, only in terms of the number and size of texts. The entire MTE multilingual comparable corpus was prepared in CES format, manually or using ad-hoc tools. The Bulgarian comparative corpus includes ***Fiction*** (texts from contemporary Bulgarian literature) and ***Newspapers*** (newspaper excerpts) subsets. The Bulgarian ***Fiction*** and ***Newspapers*** subsets were annotated manually. The data in the table below have been determined on a base of the Bulgarian fiction and Bulgarian newspapers lexica:

Part	word occurrences	distinct words	distinct MSDs in text	distinct Ctags in text
Fiction	97251	17061	313	129
Newspapers	96538	20696	295	126

The first Bulgarian electronic corpus is included in the *MTE multilingual corpus* of the MTE project (`http://nl.ijs.si/ME`), distributed on CD-ROM by *Trans-European Language Resources Infrastructure* (TELRI) Concerted Action Copernicus 1202, (`http://www.ids-mannheim.de/telri/`) for research purposes.

Bulgarian comparable corpus in Bulgarian-Polish corpus: This corpus contains approximately two million words from works of Bulgarian authors, including prose: Dimitar Talev, Dimitar Dimov, Pavel Vezhinov, Yordan Radichkov, non-fiction: Zhelyu Zhelev's „Fascism", Bulgarian translations of novels and short stories of prominent European authors.

4 Bulgarian lexical databases

CONCEDE Bulgarian LDB

The first lexical database (LDB) for Bulgarian was developed in the framework of CONCEDE project. The lexical databases of the project CONCEDE were developed on the basis of the MTE parallel multilingual corpus (so-called *Orwell* corpus). The CONCEDE project suggested a model for dictionary encoding containing a lexical database with standardized and well-understood structure and semantics. The CONCEDE project has developed lexical databases (LDBs) in a general-purpose document-interchange format for the same six MTE CEE languages: 3000-headword lexical databases for Bulgarian, Czech, Estonian, Hungarian, Romanian, and a 500-word one from the English-Slovene dictionary. The project has produced lexical resources that respect the guidelines of the Text Encoding Initiative - Dictionary Working Group (TEI-DWG), and so are compatible with other TEI-conformant resources.

The initial word lists for selection of headwords and word frequency were obtained from the MTE parallel corpus. The selection of headwords was made after word frequency and word class (POS) were taken into account, and the number of words there were in a given word-class and word-frequency band.

In order to achieve a harmonization of the LDBs according to the principal breakdown of lemmata to POS, the CONCEDE consortium decided on the following proportion: open parts of speech (nouns, verbs, adjectives, adverbs) – no more than 90 %, closed parts of speech (numerals, pronouns, conjunctions, prepositions, particles and interjections) – minimum 10% of the whole set of lemmata chosen. Under the CONCEDE project was developed an encoding scheme for lexicographic specifications of the Bulgarian language, according to the standards for electronic dictionary encoding [5]. This encoding scheme served to create the Bulgarian dictionary in the LDBs of CONCEDE. The choice of dictionary entries follows the method accepted by CONCEDE. The entries are equipped with lexicographic specifications for Bulgarian language in TEI-conformant SGML. The electronic dictionary is based on the Bulgarian Explanatory Dictionary [1]. Each entry is represented as a tree-structure. The chosen entries are divided in the following POS: noun – 33.84% of the Bulgarian sample; verb – 21.99%; adjective – 12.52%; adverb – 11.51% -- total open POS 79.86%; and numeral – 1.52%; pronoun – 5.24%; conjunction – 4.06%; preposition – 3.55%; particle – 4.40%; interjection – 1.35% -- total closed POS 20.13%. The entries in Bulgarian LDBs retain as much as possible the structure of the original paper dictionary.

The example shows the entry in the printed Bulgarian Explanatory Dictionary with the headword "**име" //***name***//:**

име *ср.* Отличително название на човек, животно и др. прен. Известност. *Той има голямо име*. грам. Категория думи, които означават предмети, качества, числа. *Съществително име. Прилагателно име. Числително име*. ◊ В името на предл. Въз основа на, заради. В името на закона. В името на свободата.

The corresponding entry in the Bulgarian LDBs:

<entry><hw>име</hw>
<gen>ср.</gen>
<struc type="Sense" n="1">
<def>Отличително название на човек, животно и др.</def></struc>
<struc type="Sense" n="2"><usg type="register">прен.</usg>
<def>Известност.</def>
<eg><q>Той има голямо име.</q></eg></struc>
<struc type="Sense" n="3"><usg type="register">грам.</usg>
<def>Категория думи, които означават предмети, качества, числа.</def>
<eg><q>Съществително име.</q></eg>
<eg><q>Прилагателно име.</q></eg>
<eg><q>Числително име.</q></eg></struc>
<struc type="Phrases">
<struc type="Phrase" n="1"><orth>В името на</orth><pos>предл.</pos>
<def>Въз основа на, заради.</def>
<eg><q>В името на закона.</q></eg>
<eg><q>В името на свободата.</q></eg></struc></struc>
</entry>

In the final phase of the development of the CONCEDE LDBs an examination was carried out – a validation process which takes two forms, "formal validation" and "content validation". The formal validation was a matter of ensuring that the databases were valid SGML documents and for the Bulgarian LDBs has been done by means of a validating SGML-parser. The content validation of the entries required human intervention and therefore was performed manually.

LDB supporting Bulgarian-Polish online dictionary

The formal model of the LDB [4] supporting the first Bulgarian-Polish dictionary is the CONCEDE model for dictionary encoding, [8]. The hierarchical structure of the dictionary entry is a tree-structure and described by 3 structural tags: **entry**, **struc**, and **alt**. The content tagset includes tags, fully describing the entry's content: the grammatical information about the headword, the translation equivalence in Polish, examples of the word's usage with translation, phrasal usage with translation (if possible) or explanation, the word's etymology (if known).

For a more adequate description of the Bulgarian verbs, two new tags are being introduced to represent the verb's conjugation (Bulgarian verbs are divided into 3 conjugations): **conjugation –** *a new tag* is added to represent the conjugation of verbs; its structure allows the subtag **type** for the possible types of conjugations of Bulgarian verbs. Furthermore, it is allowed to input additional information in the **gram** tag for the aspect – *perfect and progressive* (imperfect) of verbs, and in **subc** tag – for *transitivity/intransitivity* of verbs.

The selection of headwords included in the dictionary's LDB is based on the Bulgarian-Polish parallel corpus: the main forms (lemmata) of the most frequent word forms in the corpus are selected. The word distribution according to POS also follows the CONCEDE model: open parts of speech - no more than 90 %, closed parts of speech – minimum 10% of the whole set of lemmata chosen.

The representations of three Bulgarian verbal forms as entries in the LDB follow:

подчертава|м, -ш *vi.* podkreślać */underline/*
подчертан *part.* podkreślony */underlined/*
подчерта|я, -еш *vp.* **v. подчертавам**

```
<entry>
    <hw>подчерта'ва|м</hw>
    <pos>v</pos>
    <gram>i</gram>
    <subc>transitive</subc>
        <conjugation>
            <orth>-ш</orth>
            <type>III</type>
        </conjugation>
    <struc type="Sense" n="1">
        <trans>podkreślać</trans>
    </struc>
</entry>

<entry>
    <hw>подчерта'н </hw>
    <pos>part</pos>
    <alt>
        <pos>adi</pos>
    </alt>
    <struc type="Sense" n="1">
        <trans>podkreślony</trans>
    </struc>
</entry>
```

```
<entry>
    <hw>подчерта'|я</hw>
    <pos>v</pos>
    <gram>p</gram>
    <subc>transitive</subc>
        <conjugation>
            <orth>-еш</orth>
            <type>I</type>
        </conjugation>
    <xr>подчерта'вам</xr>
</entry>
```

Transformation of the Lexical Database to the Relational Database is carried out with the help of tables, into which the search data and indices are input. This organization allows an automatic creation of a dictionary entry for a Polish word, whenever the translation equivalence is one-to-one. Of course, the input of information about the Polish word must be done additionally.

Column / Word	**подчерта'ва|м**	**подчерта'н**	**подчерта'|я**
id	668	669	670
homonym_index			
bg_word	**подчерта#ва**	**подчерта#н**	**подчерта#**
suffix	**м**		**я**
bg_word_search	**подчертавам**	**подчертан**	**подчертая**
plural			
is_plural_rare			
conjugation	**ш**		**еш**
conjugation_type	3		1
has_gender			
gender_feminine			
gender_neuter			
id_explanation			
id_bg_word			668
referent_bg_word			**подчерта#вам**

Table bg_word

id	id_bg_word	pl_word	sense_ index	alternative_ sense_index	latin_ translation	id_explanation
1117	668	podkreślać	1	1		
1118	669	podkreślony	1	1		

Table pl_word

id_bg_word	id_characteristic
668	17
668	57
669	44
670	18
670	57

Table mm_bg_word_characteristic

id	abbreviation_ bg	abbreviation_ pl	description_ bg	description_ pl	descrip tion_ lat	id_ characteristic _ type
17	мин. нсв.	**vi**	глагол от несвършен вид			5
18	мин. св.	**vp**	глагол от свършен вид			5
44	прич	**part**	причастие			6
57	прех	**transitive**	преходен глагол			7

Table characteristic

The LDB of the Bulgarian-Polish dictionary could be used for the design and creation of new bilingual online dictionaries in the future.

5 Digital dictionaries

Monolingual: Bulgarian MTE lexica

The Bulgarian MTE lexicons (three in total) cover completely the available texts: George Orwell's novel 1984, newspaper excerpts and texts from contemporary Bugarian literature, which form Bulgarian MTE comparable corpora. Bulgarian Orwell's lexicon is a lexical list, containing 55200 entries among them 17567 lemmata, needed for use in conjunction with the morphological analyser.

The table below represents the number of lemmata and entries, distributed according to a POS-characteristic, appeared in Orwell's novel 1984:

POS	Lemmata	Entries
Nouns (total)	9891	47969
Nouns - masculine	4180	25100
Nouns - feminine	4120	16493
Nouns - neuter	1591	6376
Verbs	4140	226666
Adjectives	2155	19397
Pronouns	92	110
Adverbs	790	790
Adpositions	98	98
Conjunctions	76	76
Numerals	67	67
Interjections	172	172
Particles	86	86
Total	**17567**	**295431**

Each element of the lexicon (one entry per line) contains the following information: the inflected-form (word-form), the corresponding lemma and its standard lexical description (MSD) and has the following form:

word-form <TAB> **lemma** <TAB> **morphosyntactic description**

An excerpt from the Bulgarian lexicon follows:

Word-Form	Lemma	MSD	
бели	беля	Ncfp-n	*//nuisance, mischief; bother, trouble; difficulty//*
бели	беля	Vmia2s	*//to bleach, whiten; peel, skin; shell; hull//*
бели	беля	Vmia3s	*//to bleach, whiten; peel, skin; shell; hull//*
бели	беля	Vmip3s	*//to bleach, whiten; peel, skin; shell; hull//*
бели	беля	Vmm-2s	*//to bleach, whiten; peel, skin; shell; hull//*
бели	бял	A---p-n	*//white//*
белите	беля	Ncfp-y	*//nuisance, mischief; bother, trouble; difficulty//*
белите	бял	A---p-y	*//white//*
белия	бял	A--ms-s	*//white//*
белият	бял	A--ms-f	*//white//*
белота	=	Ncfs-n	*//whiteness//*
белота	белот	Ncms-s	*//belote (card game) //*
белота	белот	Ncmt	*//belote (card game) //*
белотата	белота	Ncfs-y	*//white, whiteness//*

Bilingual Bulgarian-Polish online dictionary

The Bulgarian-Polish online dictionary is being developed for experimental purposes. A LDB provides the language material for the dictionary. For the program realization of the web-based application the technologies Apache, MySQL, PHP and JavaScript have been used; these are free technologies originally designed for developing dynamic web pages with a lot of functionalities. The current version of the Bulgaria-Polish online dictionary works optimally with Internet Explorer 6.0+ (Windows), and with Firefox 2.0.1+ (Windows, Linux). The website resolution is 1024/768 pixels.

The web-based application consists of two primary modules: an **administrator module** and an **end-user** module.

The administrator module is intended for the person updating the dictionary, and access to it is limited only to authorized users. The administrative module is used to fill in the database and to offer user-friendly interface to the user who will be responsible for the word management. This module recognizes two types of users: (1) "**super administrator**"- who has all rights of adding, editing, deleting and searching for words; adding, editing and deleting users and (2) "**administrator**"- who has all rights except creating a new user and deleting an existing one.

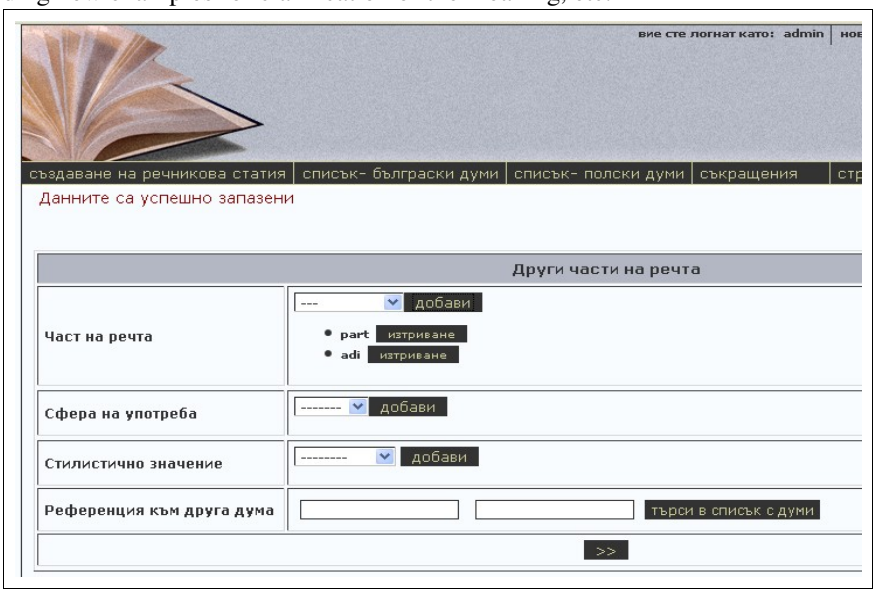
The **administrator module** manages some main sections: a section for entering a new word (see the example below), sections for searching for Bulgarian or Polish words, a section where end-users report the missing words. The Help section serves both the administrators and the end users.

Administrative panel - choosing the type of the word which will be added: **a noun**

The end-user module is the module, through which the end-user accesses the information in the dictionary. The interface is bilingual, the user can choose the input language (Bulgarian or Polish) and according to his/her choice, a virtual Bulgarian or Polish keyboard is displayed. In this way the user can choose special Bulgarian or Polish characters if they are not supported by his/her own keyboard. There are three sections in this module: a section for translating a word, an information section and a section for reporting a missing word. After making a

search for a word on the left site of the screen, a list of words is displayed starting from the given entry. A click on any of these words in the list visualizes the translation in the right frame. If we translate from Bulgarian to Polish, the whole information saved in the LDB is displayed. When translating from Polish to Bulgarian, only the Bulgarian headwords are visualized.

The program realizing the web-based application for representation of the Bulgarian-Polish online dictionary allows the dictionary volume to be expanded by adding new words, enriching the content of the dictionary entries from the LDB by adding new examples for clarification of the meaning, etc.



Administrative panel –2nd step of adding the participle

Furthermore, the structures of the LDB and of the web-based application allow a replacement of the Polish translations (texts) by texts in another language Lang. Thus, the LDB and the web-based application can be useful for the development of a new bilingual Bulgarian-Lang online dictionary.

6 Conclusion

In this paper I briefly presented the Bulgarian language resources which were developed in the Mathematical Linguistics Department at the IMI-BAS in the framework of some international projects.

Some possible directions for future work are: bringing the morphosyntactic descriptions for verbal forms in line with the Bulgarian grammar and updatiung Bulgarian MSDs and lexicon for MTE resources Version 4, extending bilingual corpora, enriching bilingual LDBs with new entries and new languages, increasing the number of headword classifiers, and increasing the speed of the search module of the web-based application for representation of an online dictionary.

Acknowledgement

I would like to thank all colleagues with whom I worked throughout the years for the development of the Bulgarian multilingual resources: Lydia Sinapova and Kiril Simov (Bulgarian Academy of Sciences, Sofia, Bulgaria), my colleagues from the MTE and CONCEDE projects, V. Koseska-Toszewa (ISS-PAS), R. Garabík (ĽŠIL-SAS), R. Panova and R. Dutsova (my students from the MSc program *Languages and Multimedia Technologies* of IMI-BAS – Veliko Tarnovo University).

References

[1] Bulgarian Explanatory Dictionary. (1997). Л. Андрейчин и др. Български тълковен речник. Четвърто издание. Допълнено и преработено от Д. Попов. Издателство Наука и изкуство, София, 1997. (In Bulgarian).

[2] Dimitrova, L., T. Erjavec, N. Ide, H. Kaalep, V. Petkevič, D. Tufiş. (1998). Multext_East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of the COLING-ACL'98*, pages 315-319, Montréal, Québec, Canada.

[3] Dimitrova, L., V. Koseska–Toszewa. (2008). Some Problems in Multilingual Digital Dictionaries. In: *International Journal Études Cognitives*. Vol. 8. SOW, Warsaw. 2008, pages 237–254. ISSN 1641-9758.

[4] Dimitrova, L., R. Panova, R. Dutsova. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Metalanguage and Encoding scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009,* pages 36–47. ISBN 978-5-9900813-6-9.

[5] Dimitrova, L., R. Pavlov, and K. Simov. (2002). The Bulgarian Dictionary in Multilungual Data Bases. *Cybernetics and Information Technologies.* Vol. 2, num. 2, pages 12–15.

[6] Dimitrova, L., R. Pavlov, K. Simov, and L. Sinapova. (2005). Bulgarian MTE Corpus – Structure and Content. *Cybernetics and Information Technologies.* Vol. 5, num. 1, pages 67–73.

[7] Dimitrova, L., P. Rashkov. (2009). A New Version for Bulgarian MULTEXT-East Morphosyntactic Specifications for Some Verbal Forms. In: Organisation and Development of Digital Lexical Resources. *Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009,* pages 30–37. ISBN 978-966-507-252-2.

[8] Erjavec, E., R. Evans, N. Ide, A. Kilgarriff. (2000). The Concede model for lexical databases. In *Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, ELRA.

[9] Ide Nancy. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 463–470, Granada, ELRA. `http://www.cs.vassar.edu/CES/`

[10] Ide N. and J. Véronis. (1994). Multext (multilingual tools and corpora). In *Proceedings of the 15th CoLing*, pages 90–96, Kyoto.

[11] EAGLES 1996. Monachini, M. and Calzolari, N. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages. EAGLES Report EAG-CLWG-MORPHSYN/R.
http://www.ilc.pi.cnr.it/EAGLES96/morphsyn/

[12] MTE 2004. MULTEXT-East Morphosyntactic Specifications – version 3, edition 10th May 2004.

Evaluating Grid Infrastructure for Natural Language Processing

Radovan Garabík[1], Jan Jona Javoršek[2], and Tomaž Erjavec[2]

[1] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
garabik@kassiopeia.juls.savba.sk
[2] Jožef Stefan Institute, Ljubljana, Slovenia
jan.javorsek@ijs.si tomaz.erjavec@ijs.si

Abstract. In this article we analyze common human language technology requirements and the possibility of implementing them using Grid infrastructure. Different possibilities for the setup of an execution environment are treated and the standard PKI based Grid security approach is explained, with an emphasis of securing data access in a potentially untrustworthy environment. Two examples of running unmodified NLP applications are presented.

1 Introduction

Increasing computing requirements for acquiring and processing large textual data-sets and working with larger and larger corpora in Natural Language Processing (NLP) and related disciplines together with ever increasing availability of computing resources allow us to work on NLP algorithms and tasks that were impractical just a few years ago. The core of the problem shifted from obtaining access to enough computation power and from optimizing algorithms into developing efficient ways of allocating the computing resources to various tasks and into finding efficient ways of dealing with huge amounts of data. Since the most accessible computing environment moved from large centralized supercomputers into the vast number of available servers connected with the ubiquitous Internet, a new paradigm in computing emerged: massively parallelized algorithms running on widely distributed networks of interconnected computers.

One of the infrastructure approaches is the Grid network, which provides a complete environment for heavy-duty computational tasks, with working solutions for user authentication, data storage, distributing load over the available resources, access control and the whole infrastructure for user management. First used mostly for computing tasks in high energy physics, the Grid is nowadays used for tasks in several different research areas. The use of Grid infrastructure for NLP has been previously discussed with several proposals (cf. [19], [4], Neuroth et al. [14], [9], [10], [12], [11], [13], reviewed in [6]).

In this article, we explore the idea of utilizing the Grid infrastructure for NLP-related tasks. Not just the computational requirements, but also commonly useful features of shared data repositories and existing Grid security features are discussed. Since the whole Grid environment runs exclusively on the GNU/Linux operating system (although there are efforts underway to port the Grid software to other Unix-like platforms, such as the BSD family and Mac OS X), our point of view will deliberately be

rather 'GNU/Linux centric', and the discussed software environment and tools will be implicitly understood to be GNU/Linux specific (unless stated otherwise).

2 Legal issues

The actual deployment of Grid computing in the natural language processing area (especially relevant for corpus linguistics) faces specific legal issues – the data being processed are in majority of cases copyrighted, and the research institutions either have very strict legal agreements governing the use of the data, or are operating entirely on copyright law sections allowing scientific and research use of the data (*fair use* in the U.S.A. jurisdiction, citation and educational use in many of the EU countries' copyright laws). The situation is somewhat similar to the problems the users of Grid computing in health care systems – though in that case, metadata are the most sensitive and protected part of the data-set, while in corpus linguistics the data (i.e. texts in the corpora) are sensitive, but the metadata is usually freely accessible [17].

In any case, the research institution using the data for research most likely does not have the right to distribute the data *at all*. If the contractual obligations prevent the institution from physically copying the data beyond the premises of the institution, it might be still advantageous to use the Grid infrastructure for computing clusters of the institution itself, and use middleware functions to restrict data-replication to those processing nodes and data storage elements physically located in the organization. This way, the whole Grid can still be used for less sensitive tasks, or for post-processing the results of operations on sensitive data (when the post-processing does not include access to sensitive data), while at the same time the computing nodes will be available as part of the whole Grid computing pool when they would be left idle otherwise.

While the actual uploading of the data to Grid-enabled storage is not to be considered a form of 'distribution' as long as no other person or organization is allowed to get the data, it is nevertheless desirable to protect the data from casual snooping. For one thing, an administrator of the Grid node where the data physically reside can get access rather trivially; and while he or she is legally obliged not to misuse his access (usually by rather strict agreements, in the case of European Grid infrastructure), a measure of additional protection seems to be necessary – to avoid data leaking in case the computer hosting the Grid node is compromised, unbeknown to the administrators. We discuss security measures used in the Grid infrastructure in the following section.

3 Software environment in the Grid

The different implementations of Grid middleware all follow the same workflow: the user has to provide a way to parallelize the computing task in a number of jobs that can be run in parallel and encode the solution by providing a *job specification*, indicating the data files required (using URIs), suitable computing environment (ABI, API, execution environment), computing time and resources needed (wall-time, RAM, disk) and a way to store the results.

This can be done manually, and a job is then submitted with a dedicated command, or, alternatively, can be done with the help of a dedicated web application, usually domain or experiment specific.

The system then takes care of selecting the appropriate Grid site and free worker nodes, downloading the data files, making the pre-installed software (execution environment) available and starting the job script. A number of tools enable the user to monitor the progression of the task, including its standard input and standard/error output, working directory contents etc.

When the job finishes, it may upload the resulting data files to a Grid-enabled disk storage, or, alternatively, the user can use command-line tools to download the data from the working directory directly.[3]

To make the system work reliably and securely, a number of information, monitoring and accounting systems are part of the infrastructure. In addition, an advanced security model is used to ensure resource protection and data integrity. As we have to consider this system's suitability for protection of copyrighted data-sets in linguistic resources, a short overview is presented in the following subsection.

3.1 Security

Computing grids had to be very security-conscious from the very beginning, since the very premise of a Grid network is, from the point of view of the site administrator, to give external users access to the local computing infrastructure and, from the point of view of Grid users, to entrust data and applications to untrusted, foreign sites.

Moreover, the basic requirement for a viable, scalable and sustainable security infrastructure in the context of large Grid networks has to be a robust solution with as few single points of failure as possible to avoid failures of security services that could effect negatively the availability of the whole infrastructure.[4]

Grid security has several components: **(a)** Authentication, a method of confirming the identity of the user or organization behind an operation, is implemented on the basis of the Public Key Infrastructure (PKI) and standard x509 digital certificates (with a number of extensions to facilitate the use of PKI in the context of Grids). **(b)** Authorization is provided in the framework of virtual organizations (VOs), a mechanism enabling Grid users all over the world to organize themselves according to research topics and computing requirements, regardless of geographic constraints, and permitting sites to regulate the use of their resources according to user, discipline, software requirements etc. **(c)** Monitoring and ticketing permits users and administrators to keep track of infrastructure availability and to react to technical and security matters in a timely fashion. **(d)** Accounting reports on the use of the infrastructure and enables the community to regulate and enforce the use of the infrastructure.

Public Key Infrastructure. Public Key Infrastructure, first introduced to the general public in the context of securing the web and enabling on-line shopping and banking, has become the standard authentication model in many application domains. Defined

[3] See [5] for architecture overview of the NorduGrid ARC middleware.

[4] For an overview, see [8].

by a number of Internet Drafts, RFCs and standards, PKI is a widely deployed and evolving system.[5]

PKI is based on the property of asymmetric ciphers, where a different key is used for encryption and decryption. This property allows the encryption key to be always kept private and secret and the decryption key to be public, usually published with some information about the owner of secret key in the form of a x509 digital certificate.

In PKI, such a digital certificate is used as the token of identification: it is issued by a certification agency (CA) on the basis of an identification process (i.e. checking legally acceptable personal ID documents in person). But the certificate is coupled with a secret key that has been generated by the user requesting the certificate and is never exposed to the CA. To issue a certificate, the CA now sets up information about the entity (user, host or service) to be certified in accordance to the identification data provided in a standard form called a Distinguished Name (DN, following a LDAP-like name scheme: `CN = Joe User, OU = My Department, DO = Institute of Dispersive Linguistics, DC = San Marino`, and signs it with their own secret key from the CA certificate.

This scheme ensures that nobody, not even the CA, can use the certificate (since only the owner of the certificate possesses the secret key) and protects the information in the certificate with the signature, produced with the CA's own secret key.

To make the system work, CA certificates with public keys are published in a well advertised manner (or shipped with software, such as. web browsers, Grid middleware packages and GNU/Linux distributions). Recipient of a document or a connection that uses a client certificate and is encrypted or signed with such a certificate can therefore verify that the document or connection really was encrypted or signed by the said certificate by decrypting it with the public key included in the certificate, and it can verify the information in the certificate by checking the certificate with the CA public key in the same manner.

A number of additional security measures are used in the Grid: CA secret keys are kept in off-line systems or in dedicated certified hardware modules (hardware security modules or HSM) while end-entity certificates are re-issued with new keys yearly or kept in hardware security tokens. In addition, actual user certificates are never entrusted to non-trusted entities: for almost all operations in the Grid, short-lived proxy certificates are used instead (described bellow).

Virtual organizations. While PKI provides authentication, a different system is needed to provide authorization, i.e. to help decide if a given user, host or service is to be allowed to carry out a specific task: use a specific resource or access specific data. In the context of Grid computing infrastructure, this role is implemented in the framework of virtual organizations (VOs).

A Virtual Organization serves two purposes: **(a)** As an organizational form, a VO permits a number of researches from different organizations, usually geographically dispersed, to collaborate and share tools, data and resources. **(b)** In the Grid security infrastructure, a VO provides means of regulating access to resources, i.e., a VO provides authorization after authentication is provided by PKI.

[5] See [2] for an extensive up-to-date overview of the relevant documents.

With this combination of roles, Virtual Organizations have proven themselves to be most efficient in enabling a higher level of international collaboration and have permitted the European Grid network to foster new, faster development in many disciplines by providing an unprecedented framework for international collaboration.

In practice, members of a research project or a discipline can set up a VO and decide on its modes of operations and access to resources quite independently. They have to decide what kind of tools the VO members will be using in the Grid, define the data formats, prepare data repositories, develop execution environments with the tools installed and set up a Virtual Organization Membership Service server (VOMS server) to store authorization credentials.

Then some resources have to be made available to the community of VO members. In practice, that means obtaining support of a number of Grid sites (organizations owning computing clusters partaking in the Grid) that have to configure their Grid middleware installations to include the new VOMS server in its authorization procedures and to either install the execution environment (or, more realistically, environments) for the VO or give access to some members of the VO so that they can perform the installation and maintenance if the execution environment on the site themselves. Additionally, a number of Grid storage elements (SE) has to be configured to allow the VO members to access and store the data on their disk space.

Proxy certificates. With the VO and VO supporting Grid sites, a VO member can submit Grid jobs and access VO-owned data using his certificate. This is implemented in an indirect manner by means of Grid proxy certificates, as mentioned previously in the discussion of PKI infrastructure.

Grid proxy certificates are **(a)** primarily used to permit a job to authenticate in the name of the user spawning the job, without the requirement of direct user interactions during the course of the job. This means that the proxy certificate must have the same DN as the users' certificate, but it has a different secret key which is not protected with a pass-phrase that would require user interaction on the keyboard. Proxy certificates are generated with a tool that uses the users' certificate to sign the proxy (as if it were a CA), thus confirming that the proxy was indeed generated by the user. In addition, gird proxy certificates are protected with file permissions and are always short-lived (from several hours to a few weeks) to mitigate the risk of the unprotected secret key.

To interact with the VO authorization system, the user generates a VOMS Grid proxy certificate that obtains special certificate extensions from the VOMS server and incorporate them in the proxy certificate. These extensions encode VO group and role attributes of the user and are themselves signed by the VOMS server with its service certificate, using the PKI infrastructure's authentication facilities to implement an authorization layer.

In this manner, a job can obtain authorization to use computing resources and data simply by providing a suitable VOMS proxy certificate. Its attributes are recognized by the Grid manager servers that provide it with to data storage (storage resource managers, SRM) and other resources.

As and additional level of security, Grid managers assign each job a temporarily unique user ID in the underlying operating system mapped from its active VO role in

such a way that no jobs with different roles (and therefore potentially different access permissions) can share access on the underlying implementation.

In this way the system implements fine-grained control over the use of Grid resources and data without any reliance on the availability of authentication and authorization servers, thus avoiding a single point of failure that would have a significant impact on the scalability of the system.

Data Protection. Using these security components, additional measures of data protection can be implemented when necessary. In the context of NLP, such a measure is of critical importance, since most of the data-sets in corpus linguistics contain copyrighted texts that need to be protected.

To solve this problem, the corpus data has to be suitably protected where it is permanently stored. Therefore we propose to store the corpus data in encrypted form in a dedicated storage element and set up the access authorization in such a way that access is restricted to VO users who belong in a VO group of users who signed the necessary legal agreements to access the data. Furthermore, we propose that the data is transferred to the untrusted environment of Grid worker nodes, where jobs perform their computations, in the encrypted form and that the decryption keys are issued to the jobs protected with asymmetric encryption decryptable only by the job's Grid proxy keys so that only the jobs can access the keys and decrypt the data.

In this manner, access and decryption is regulated with the authorization of embedded VOMS attributes in the proxy certificate without any additional authorization steps, while the data is never shipped or stored in unencrypted form.

If the tools used by the job have to store temporary files on disk, these are protected from other processes (with the exception of system administrators, who are already bound by strong agreements pertaining to data security on the Grid) and are in addition of short-lived nature.

There exist different implementations of the system described. The simplest form involves the use of a decryption filter in the job script and is rather simple to deploy. A more flexible solution, based on CryptoSRM (cryptographic storage resource manager) and Hydra Key Storage (a distributed fragmented encryption key storage system) is described in [17].

4 General requirements of NLP related tasks

Contemporary NLP tasks are rather varied; some of them require a lot of "pure" computing power, but many tasks, especially in the area of corpus linguistics, merely process large data files. From the software point of view, the tools used cannot be more diverse – they are often programmed in typical computer languages, like C or C++, but a lot of data processing is done in scripting languages, such as Perl or Python, and Java is increasingly popular, and more often than not, one specific task uses several different tools bound by short programs written in a shell script. The use of (high level) scripting languages even for the computing intensive tasks means that the analysis is less effective than it could be, but the ease of creating and maintaining the tools more than outweights this particular disadvantage. From this follows than the tools are often

fragile and require a specific environment, which sometimes means that even using a different GNU/Linux distribution that the one the software has been developed on can be a major problem.

The Grid environment, due to its initial connection with the use in High Energy Physics, predominately uses Scientific Linux CERN distribution (SLC) version 4 for the job computing environment (with a changeover to version 5 currently in progress). The ideal solution would be of course to put all the necessary NLP software into the execution environment (which is available at each of the computing nodes) and use the standard distribution. It is, however, sometimes much more convenient to use an operating system environment more suitable for the users and their tools. There are two possible solutions: to run under a chroot environment or to use virtualization. Both options are discussed below.

4.1 Userspace and full virtualization

Chroot is a UNIX system that changes the effective root of the filesystem for the process and its children. The basic usage for chroot is twofold: it can be used to restrict untrusted (or potentially dangerous) processes from accessing the rest of the filesystem, or it can be used to run processes in a different filesystem environment (different filesystem layout with different system executables and dynamic libraries). It should be noted that chroot does not offer true virtualization since isolation from the host system is not complete – in particular, system kernel, networking subsystem and process management are shared with the host system, so that the processes in the chroot environment cannot bind to sockets that are used on the host system (and vice versa), and if process management is to be possible in a chroot environment, the `proc` filesystem has to be mounted inside chroot environment, enabling the guest to access the information about host processes[6].

On the other side of the spectrum, there are complete virtualization solutions, emulating the guest system. These can emulate the CPU completely in software (approach commonly used in emulating vintage computers on modern operating systems, or when a computer platform switches the architecture), or run the guest machine natively, trapping and emulating only privileged or unimplemented instructions. Modern computer architectures usually offer dedicated hardware features to facilitate the implementation of virtual machines[7], some mainframe architectures even offer complete, seamless virtualization in hardware.

Then there are several different approaches that lie somewhere in between those two extremes, ranging from **(a)** paravirtualization, which requires cooperation from the guest operating system kernel (in order to achieve negligible performance loss due to the virtualization), used e.g. by the `XEN` virtualization solution; to **(b)** compartmentalization (i.e Linux virtual servers and `OpenVZ`), which divides the host operating system into different compartments with completely separated processes, network access and

[6] This does not matter as much as it seems as long as the chrooted processed run under different PID from the host ones, because a non-root user cannot affect other processes, and a chrooted superuser can break out of the chroot anyway.

[7] Until rather recently (before the introduction of VT-x and AMD-V), such features were not available in common Intel-compatible off-the-shelf computers.

filesystems but sharing the same kernel; to vanilla kernel namespace support, which only separates user and process management (slightly extending chroot separation).

The virtualization techniques mentioned differ on performance impact [15] – ranging from none at all in case of a simple chroot or chroot with namespaces, over very little for `OpenVZ`-like compartmentalization to a more significant one for full virtualization. The specific areas of impact vary, too – while the raw CPU performance rarely decreases by more than a few percent (with the exception of complete software emulation of the guest architecture), I/O penalties are sometimes severe.

From our point of view, the best way to use the specific software is to install it inside a runtime environment which is made available to the jobs when submitted to the Grid. This is directly supported by the Grid infrastructure and requires no additional steps or privileges. However, at this time this requires a significant effort, since all the tools and their dependencies have to be compiled (or installed in a non-standard location inside the runtime environemnt) on the standard SLC distribution, which can be a problematic if the software has many external dependencies.

Installing a chroot environment, on the other hand, enables us to avoid porting the software to the SLC distribution – inside the chroot, we can install any reasonably standard GNU/Linux distribution and any necessary software packages. In addition, many of the commonly used distributions already have support for (at least partial) installation inside a chroot environment built in. But in the context of Grid infrastructure this solution has a significant disadvantage, since it requires support from the cluster administrator since chroot environments are not a standard feature of the Grid environment.

Using a complete virtual machine allows us to run a complete GNU/Linux distribution, with completely separate networking support and user management, including the ability to run processes with superuser privileges, and the ability to use filesystems otherwise not supported by the host system[8]. But the main advantage is the possibility to run completely different operating system[9]. However, installing and using virtual machines requires not just administrator cooperation, but often also nonstandard host operating system extensions (such as special kernel modules). One of the more interesting virtualization systems in this context is User Mode Linux, which does *not* require any special host support, runs as an ordinary user process and provides a complete guest Linux kernel environment. Unfortunately, guest environment in this case suffers from a big I/O performance degradation, which can be a noticeable problem when dealing with very large corpus data.

While there is significant research in the use of different kinds of virtualization in the context of Grid technologies, this is not a wide spread feature at this time. We have been able to use clusters with full support for chroot environemnts, but we realize that for quick adoption and widespread use of Grid computing in NLP, porting of tools to the most often supported environment, i.e. SLC, will be necessary.

[8] Such as encrypted filesystems.

[9] Therefore we can use e.g. the tools available only for Microsoft® Windows® family of operating system, if we can get around their mostly point-and-click nature and run them noninteractively.

5 Proof of concept

This section presents and experiment in using the Grid to execute two NLP tasks for the Slovak language. The first subsection introduces TectoMT, a machine translation system, the second morphosyntactic tagging with morče, and the third gives the experimental usage of the two systems on the Grid.

5.1 TectoMT

TectoMT [20] is a software framework aimed at machine translation at the tectogrammatical level of analysis. The system is modular – the framework itself consists of many independent modules (*blocks* in TectoMT terminology), each implementing one specific, independent NLP-related task. Each of the blocks is a Perl module that interacts with the system using a single, uniform interface. However, sometimes the module serves only as a wrapper for the underlying implementation in another programming language. The tectogrammatical annotation and consequently the TectoMT framework primarily stores linguistic data in its own format, called TMT. TMT is an XML-based format, designed as a schema of the Prague Markup Language (PML)[10] [16]. Nevertheless, its blocks are by no means obliged to use this format.

TectoMT has been developed with modern Linux systems in mind, and as such its installation requirements are easily met by any contemporary Linux distribution. It should be noted that TectoMT, being written mostly in Perl, depends on many external Perl modules and its installation scripts are intelligent enough to automatically download and install any missing dependencies; this, however, circumvents standard distribution packaging systems, therefore it is better to install all the necessary packages with the packaging system tools before attempting to install TectoMT. There are also some C language modules that are not compiled by default, but have to be compiled separately inside the TectoMT installation source tree.

TectoMT also has some built-in capabilities for parallelization of its tasks, using the Sun Grid Engine – it is possible to adapt the Sun Grid Engine batch software to various Grid middlewares [3], but TectoMT can be run on the Grid system directly without relying on its internal parallelization possibilities, if the user takes care of splitting the input data into appropriate chunks for parallel processing.

5.2 Tagging a corpus

Morphosyntactic tagging of the Slovak National Corpus consists of two steps. The first performs morphosyntactic analysis, where each word in the input texts is assigned a set of possible morphosyntactic tags. This step essentially consists of looking up the possibilities of lemma/tag combinations in a constant database table using the wordform as a key, with an additional step for unknown words, where the list of possible tags is derived from the similarities of word endings to the ones present in the database tables. The software is implemented in the Python programming language and is actually quite fast, since the core of the task consists simply of a look-up in the possibilities in the

[10] Not to be confused with the Physical Markup Language.

tables, and most of the CPU work is spent on I/O operations, parsing the input file and assembling the output. On a reasonably recent hardware (Intel Xeon 2.33 GHz CPU) it is able to process over 10 000 words per second. It can also parallelize easily, since the words can be analyzed independently of each other.

The second step is disambiguation, where each word is assigned a unique lemma and a morphosyntactic tag out of the possibilities assigned in the first step. For disambiguation, we use *morče*, an averaged perceptron model originally used for the Czech language tagging [18], re-trained on the Slovak manually annotated corpus. Disambiguation is much slower that the morphology analysis, its average speed reaches only about 300 words per second. Parallelization at the application level is also not possible without some redesign of the *morče* itself, but the nature of tagging makes it easy to split the input data into as many chunks as we want and run *morče* instantiations in parallel.

Given the speed differences between morphology analysis and disambiguation, we can safely consider the morphology analysis execution time negligible and design the whole tagging to be done in one step, without the need to parallelize the morphology analysis process while the disambiguation is to be run in parallel.

5.3 Installation and usage

As our GNU/Linux distribution of choice is Debian[11], we did all the testing on the Squeeze (testing) Debian distribution, which is a "moving target" distribution, meant for users that want newer version of the distribution and included packages, but do not want to deal with (potentially) broken bleeding edge packages from the *unstable* Debian repositories. To summarize, a package will get into *testing* if it has no release-critical bugs, has spent several days in the *unstable* repository and its inclusion in *testing* will not break other packages. We used *testing* deliberately, because it is advantageous to use new versions of the required packages which will not become obsolete in near future, even if the packages in *testing* repositories will be rather quickly replaced by still newer versions.

Debian has a standard method for installing the base system into a chroot environment, implemented by a tool, called `debootstrap`[1].

Once the chroot environment was created, we proceeded and installed TectoMT Perl dependencies into the chroot system. Finally, we ran the TectoMT installer. Similarly we installed the Slovak version of the *morče* software from within the chroot. The installation process was unremarkable comparing to a regular installation.

After the chroot environment had been prepared and installed on the Grid servers, we were able to submit batch jobs with either TectoMT or the disambiguation tasks. We used the Grid infrastructure to morphosyntactically tag and analyse the spoken Slovak language corpus [7]. We divided the corpus (434 676 words) into 11 approximately equally sized chunks and submitted them to be tagged in 11 independent batch jobs, then joined the results. The run times ranged from 103 to 236 seconds, with the average time around 141 seconds per job[12]. Even including the overhead spent on submitting

[11] `http://www.debian.org`

[12] Times mentioned are *wall times*, i.e. total time elapsed from the beginning to the end of the task, not including the time spent waiting in the job queue or downloading the data files.

and waiting for the job to start, this is a significant reduction of the total time needed to tag the corpus. In fact, as far as the typical Grid tasks are concerned, our jobs were very short, and we could achieve less overhead by running longer jobs (a typical Grid jobs takes from 2 hours to more than a day). Such a s setup will actually be typical when tagging a bigger, representative text corpus.

6 Conclusion

We demonstrated the possibility to use the Grid infrastructure for NLP related task. From user point of view, a Grid computer behaves like any ordinary workstation running Scientific Linux CERN distribution; however, in this article we discussed different methods of using a custom GNU/Linux environment (or even another operating system) to better support the tools needed by the user. We set up and tested a chroot environment running Debian GNU/Linux *Squeeze* distribution. Installing and running TectoMT framework inside the chrooted environment was straightforward; similarly we experienced no obstacles in installing and using the Slovak language morphology tagger – we therefore do not expect any problems in deploying all kinds of "well behaved" Unix (Linux) based software.

However, in order to truly exploit the Grid potential, we envisage a scheme where the linguistic data (especially text corpora) are stored on the Grid infrastructure as well, and the existing Grid access control infrastructure is extended in order to be provide secure access to the data to third parties interested in accessing the data in such a way that all the limitations and conditions arising from the copyright law and other binding agreements are met.

In this way, we hope that the Grid infrastructure will soon become available to researches in computational linguistics and, by multiplying the computing resources available, will speed up linguistic research tasks and enable us to develop new algorithms, research methods and tools.

References

[1] Installing new Debian systems with debootstrap. `http://www.debian-administration.org/articles/426`. Retrieved 2009-10-05.

[2] Public-Key Infrastructure (X.509) (pkix). `http://www.ietf.org/dyn/wg/charter/pkix-charter.html`. Retrieved 2009-10-10.

[3] Borges, G., David, M., Gomes, J., Fernandez, C., Lopez Cacheiro, J., Rey Mayo, P., Simon Garcia, A., Kant, D., & Sephton, K. (2007). Sun Grid Engine, a new scheduler for EGEE middleware. In *IBERGRID – Iberian Grid Infrastructure Conference*.

[4] Carroll, J., Evans, R., & Klein, E. (2005). Supporting Text Mining for e-Science: the challenges for Grid-enabled Natural Language Processing. In *Proceedings of the UK e-Science All Hands Meeting*.

[5] Ellert, M., Gronager, M., Konstantinov, A., Konya, B., Lindemann, J., Livenson, I., Nielsen, J., Niinimaki, M., Smirnova, O., & Waananen, A. (2007). Advanced Resource Connector middleware for lightweight computational Grids. *Future Generation Computer Systems*, *23*(2), 219–240.

[6] Erjavec, T. & Javoršek, J. J. (2008). Grid Infrastructure Requirements for Supporting Research Activities in Digital Lexicography. In *Mondilex: Lexicographic Tools and Techniques*, (pp. 5–13). IITP RAS.

[7] Garabík, R. & Rusko, M. (2007). Corpus of Spoken Slovak Language. In Levická, J. & Garabík, R. (Eds.), *Computer Treatment of Slavic and East European Languages*, Brno. Tribun.

[8] Laccetti, G. & Schmid, G. (2007). A framework model for grid security. *Future Generation Computer Systems*, *23*(5), 702–713.

[9] Luís, T., Martins de Matos, D., Paulo, S., & Ribeiro, R. D. (2008). Natural Language Engineering on a Computational Grid (NLE-GRID) T5 – Performance Experiments. Technical Report 35 / 2008, INESC-ID, Lisboa.

[10] Martins de Matos, D., Luís, T., & Ribeiro, R. D. (2008). Natural Language Engineering on a Computational Grid (NLE-GRID) T1 – Architectural Model. Technical Report 38 / 2008, INESC-ID, Lisboa.

[11] Martins de Matos, D. & Ribeiro, R. D. (2008). Natural Language Engineering on a Computational Grid (NLE-GRID) T2h – Encapsulation of Reusable Components: Lexicon Repository and Server, January 2008. Technical Report 32 / 2008, INESC-ID, Lisboa.

[12] Martins de Matos, D., Ribeiro, R. D., Paulo, S., Batista, F., Coheur, L., & Pardal, J. P. (2008). Natural Language Engineering on a Computational Grid (NLE-GRID) T2 – Encapsulation of Reusable Components. Technical Report 31 / 2008, INESC-ID, Lisboa.

[13] Marujo, L., Lin, W., & Martins de Matos, D. (2008). Natural Language Engineering on a Computational Grid (NLE-GRID) T3 – Multi-Component Application Builder. Technical Report 33 / 2008, INESC-ID, Lisboa.

[14] Neuroth, H., Kerzel, M., & Gentzsch, W. (Eds.). (2007). *German Grid Initiative D-Grid*. Universitätsverlag Göttingen.

[15] Padala, P., Zhu, X., Wang, Z., Singhal, S., & Shin, K. G. (2007). Performance evaluation of virtualization technologies for server consolidation. Technical report, HP Laboratories.

[16] Pajas, P. & Štěpánek, J. (2006). XML-based representation of multi-layered annotation in the PDT 2.0. In Hinrichs, R. E., Ide, N., Palmer, M., & Pustejovsky, J. (Eds.), *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, (pp. 40–47)., Genova, Italy.

[17] Santos, N. & Koblitz, B. (2008). Security in distributed metadata catalogues. *Concurrency and Computation: Practice and Experience*, *20*(17), 1995–2007.

[18] Spoustová, D., Hajič, J., Raab, J., & Spousta, M. (2009). Semi-supervised training for the averaged perceptron POS tagger. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 763–771)., Morristown, NJ, USA. Association for Computational Linguistics.

[19] Tamburini, F. (2004). Building distributed language resources by grid computing. In *Proc. of the 4th International Language Resources and Evaluation Conference*, (pp. 1217–1220).

[20] Žabokrtský, Z., Ptáček, J., & Pajas, P. (2008). TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, (pp. 167–170)., Columbus, OH, USA. Association for Computational Linguistics.

Synset Building Based on Online Resources

Ján Genči

Department of Computers and Informatics
Technical University of Košice, Slovakia

Abstract. A WordNet and EuroWordNet are two well-known projects the aim of which is to build so-called synsets (set of synonyms) in order to document meanings of the words. The purpose of the WordNet project was to build English database. The purpose of the EuroWordNet project was to develop "national" versions of WordNet databases including so-called "interlingua index" for seven European languages. Since Slovak language was not included in the original EuroWordNet project, there is an initiative to find the way how to build relevant Slovak synsets and connect them to the Interlingua Index. The paper describes several approaches which can be used to build Slovak synset equivalents using (mainly) on-line resources (resources available on the Internet).

1 Introduction

A WordNet [1] and EuroWordNet [2] are two well-known projects the aim of which is to build so-called synsets (set of synonyms) in order to document meaning of the words. WordNet specifies semantic network of English words. Project ontology for particular sensets of words defines (beside others):

- More specific senses (hyponyms)
- More general senses (heperonyms)
- Senses of which semantic entity consists (meronyms)
- Explanation of sense by:
    - set of synonyms or simple word
    - description
    - example

Presented approach means standard way of sense specification in mono- and bilingual dictionaries. The purpose of the WordNet project was to build English database. The EuroWordNet [2] project developed "national" versions of WordNet databases including so-called "interlingua index" for seven European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The Interlingua Index links the EuroWordNet synsets to the original WordNet synsets.

WordNet and EuroWordNet proved themselves as valuable source of information not only for linguists but also for computer scientists working in the area of text processing. Computer linguistics can use both networks of senses and senses itself (specified as set of synonyms) in the process of solving different tasks. That is a reason why scientists from many other countries are interested in having their "national" version of WordNet. As an example we can state BalkaNet project, purpose of which was to "enrich" EuroWordNet database by some other languages (Greek, Turkish, Romanian, Bulgarian and Serbian) [3].

Since Slovak language was not included in the original EuroWordNet project and there were no (mainly financial) sources to build Slovak compilation of WordNet, there is question if it is possible to automate process of finding relevant Slovak synsets and connect them to the synsets in other languages and moreover link then to Interlingua Index.

2 Sources of synonyms (synsets)

The sets of synonyms for particular language can be acquired by several ways. The first, "direct" approach is to use sources which specify synonyms explicitly. Other ways, "indirect", are based on algorithms for synset extraction from available sources. They can use either explicit synset specification or combination of several sources.

2.1 "Direct" sources of synsets

As explicit sources of synsets can be regarded:

- WordNet [1] – English synsets
- EuroWoirdNet [2] – Dutch, Italian, Spanish, German, French, Czech and Estonian
- BalkaNet [3] – Greek, Turkish, Romanian, Bulgarian and Serbian
- Thesauri for particular languages

All resources stated above contain synset specified in the direct way. As another, quite straightforward source of synsets we can regard monolingual or translation dictionaries (bilingual or multilingual). Monolingual dictionaries specify senses of words in a specific way usually as list of senses. Translation dictionaries use notation very similar to notation used in the WordNet (or vice versa) – list of synonyms, description of sense and examples. Public availability of presented sources varied. While WordNet and on-line translation dictionaries are publicly available, EuroWordNet and BalkaNet are closed project and they are available by special permission only.

2.2 "Indirect" sources of synsets

Series of synonyms can be found by many different ways using various sources. While direct sources of synsets can be used as source of synset for particular language at the same time they can serve as an indirect source for synsets for other languages. That was a case of the WordNet, which was used as source of synsets in the EuroWordnet project.

Translation dictionaries we regard as direct source of synsets. However, on-line translation dictionaries represent special case. They usually do not specify translation of particular senses of the word but state the list of all translations only. That is a reason, why we regard them as indirect sources.

The paper presents several ways of producing Slovak synsets linked to synsets in other languages. We present approaches based on the WordNet, and on-line translation dictionaries. More complex approach is also mentioned.

3 Formal model of synset translation

It is clear that translation of the words between languages have to be based on the senses of the word. We try to formalize this process for better understanding.

Def. Let W_i is a word, expressing sense V. Synset S (set of synonyms) is a set of words, expressing sense V. Formally:

$$S=\{W_i \mid 0<i<n\},$$

where n is a number of synset members.

Def. Let V_i is a sense, expressed by a word W. Homset H (set of homonyms) is a set of all senses expressed by word W. Formally:

$$H=\{V_i \mid 0<i<n\},$$

where n is a number of senses.

Let us take word W. Generally, word W can have several senses, what means it is a member of homset H, with r members. Every sense of homeset H, can be expressed by synset S_i ($1 \le i \le r$) (Fig. 1).

$$\begin{aligned} S_1 &= \{w_{11}, w_{12}, \ldots, W, \ldots, w_{1s_1}\} \\ S_2 &= \{w_{21}, w_{22}, \ldots, W, \ldots, w_{2s_2}\} \\ &\ldots \\ S_r &= \{w_{r1}, w_{r2}, \ldots, W, \ldots, w_{rs_r}\} \end{aligned}$$

Fig. 1. Senses of word W specified by corresponding synsets

Example

To illustrate presented approach we use on-line version of WordNet [1]. Let us take word accident. Result provided by the WordNet is presented on the Fig. 2.

We can see that homset (word) accident has two members (i.e. has two senses):

1. an unfortunate mishap; especially one causing damage or injury
2. anything that happens suddenly or by chance without an apparent cause

The first one is a synset consisting of a single word, second one is consisting of several words (accident, stroke, fortuity, chance event).

Def. Let $T_{L_1}^{L_2}(W)$ is an operation of translation of word W from language L_1 to language L_2.

Translation $T_{L_1}^{L_2}(W)$ can be semantically expressed according Fig. 3.

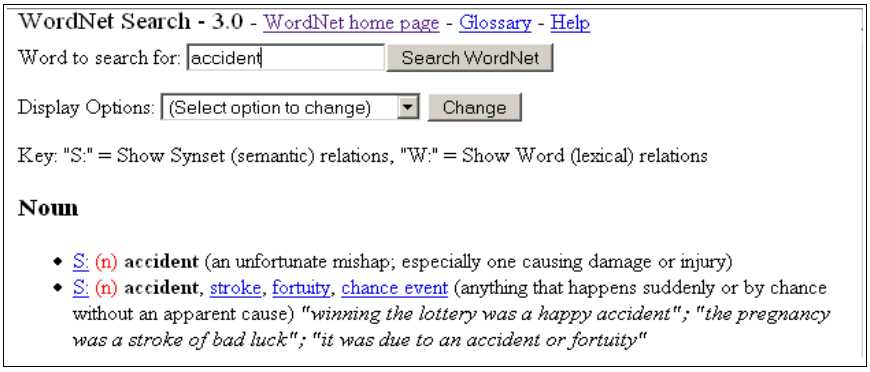


Fig. 2. WordNet result for word accident

$$S_1^{L_1}=\{w_{11}^{L_1}, w_{12}^{L_1}, \ldots, W^{L_1}, \ldots, w_{1s_1}^{L_1}\} \rightarrow V_1 \rightarrow S_1^{L_2}=\{w_{11}^{L_2}, w_{12}^{L_2}, \ldots, w_{1i}^{L_1}, \ldots, w_{1t_1}^{L_2}\}$$
$$S_2^{L_1}=\{w_{21}^{L_1}, w_{22}^{L_1}, \ldots, W^{L_1}, \ldots, w_{2s_2}^{L_1}\} \rightarrow V_2 \rightarrow S_2^{L_2}=\{w_{21}^{L_2}, w_{22}^{L_2}, \ldots, w_{2i}^{L_1}, \ldots, w_{2t_2}^{L_2}\}$$
$$\ldots$$
$$S_r^{L_1}=\{w_{r1}^{L_1}, w_{r2}^{L_1}, \ldots, W^{L_1}, \ldots, w_{rs_r}^{L_1}\} \rightarrow V_r \rightarrow S_r^{L_2}=\{w_{r1}^{L_2}, w_{r2}^{L_2}, \ldots, w_{ri}^{L_1}, \ldots, w_{rt_r}^{L_2}\}$$

Fig. 3. Semantics of translation of word W

The result of the operation $T_{L_1}^{L_2}(W)$ we can express as $\{S_1^{L_2}, S_2^{L_2}, \ldots, S_r^{L_2}\}$

It is clear that word *W* is member of each synset $S_i^{L_1}$. Every sense represented by synset in the source language is translated to destination language as a set of words – synset which represents translated sense in the destination language.

Using two-way translation (forward and back) for given word *W* we can try to „reconstruct“ particular synsets using translation dictionary according to following algorithm:

1. $get\ T_{L_1}^{L_2}(W)=\{S_1^{L_2}, S_2^{L_2}, \ldots, S_r^{L_2}\}$
2. $for\ each\ w_i \in S_i^{L_2}$
   a) $R_i=T_{L_2}^{L_1}(w_i)=\{S_1^{L_1}, S_2^{L_1}, \ldots, S_q^{L_1}\}$ *// get back translation*
   b) $S_i^{L_1}=\bigcap_{k=1}^{t_i} R_k$

Presented algorithm gives us possibility to build relevant (linked) synsets in both languages.

4 Synset building – theoretical view

On the base of presented model we have implemented or plan to implement several approaches to build synsets for Slovak language:

1. using synsets from the WordNet database
2. using bilingual on-line translation dictionaries
3. using (publicly) available thesauri

4.1 Using WordNet database

As stated above the WordNet specifies senses of words by synsets. Moreover, it specifies relationships between word senses – more general and more specific word senses. To determine synsets of destination language (Slovak language in our case) we are using two approaches:

1. intersection of translation of particular WordNet synsets words
2. translation of related words in the WordNet network (hyponyms and hyperonyms)

The first approach we use in the case if particular synset is defined by multiword synset. The second one is used in the case of single word synset.

We try to illustrate our approach on the case of word computer. According to WordNet, this word has two noun senses:

1. computer, computing machine, computing device, data processor, electronic computer, information processing system (a machine for performing calculations automatically);
2. calculator, reckoner, figurer, estimator, computer (an expert at calculation (or at operating calculating machines));

Result of determination of Slovak equivalent by intersection of translation of English synset members is presented in the Fig. 4.

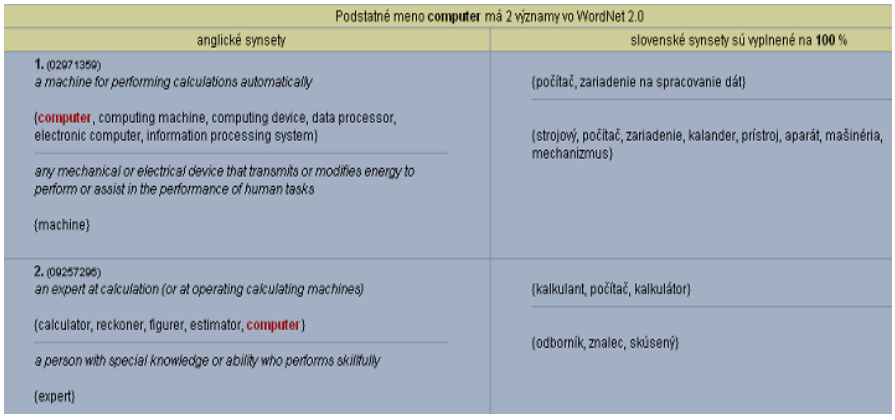


Fig. 4. Building Slovak synset based on WordNet data

Using specified approach for several languages it is possible to build linked multilanguage synsets.

4.2 Using on-line translation dictionaries

Classical printed dictionaries specify senses of word as lists of words in the destination language. For many reasons this lists are not "full" synsets. However, even not complete synsets can be used to determine the translated sense of word using similar approach we used in the WordNet case. The only problem is on-line availability of such source.

In the contrary to printed version of dictionaries, corresponding on-line dictionaries usually lack information regarding sense specification. For example, from the first ten results of Google search condition "slovnik site:sk", the only one dictionary [4] provides results in the structured form. Other ones provide only unstructured translation of given word. However, even such approach gives us opportunity, using forward and back translation, to determine some synsets in both languages.

Let us take Slovak word *jednotka*. Translating this word from Slovak to English using dictionary FEEHA [5] we got translation specified by tab.1. (results for multiword senses are eliminated). Intersecting backward translations (with simultaneous elimination of original word jednotka) we got result according Tab. 2. Adding word *jednotka* to computed synsets we get results according Tab 3.

Translation to English	Backward translation to Slovak
drive	jazda, pohon, nápor, vychádzka, cesta, vozovka, budenie, jednotka, mechanika, zariadenie, podnikavosť, motivácia, prevod, nábor, riadenie, výjazd
element	člen, častica, živel, súčiastka, časť, článok, prvok, element, jednotka, zložka
entity	podstata, jednotka, celok, bytosť, bytie, teleso, objekt, entita, existencia, predmet
force	sila, úsilie, vplyv, moc, vodopád, tlak, tiaž, zbor, oddiel, jednotka, platnosť, účinnosť, pôsobnosť, násilie, donútenie, zmysel, tvárnik, zasahovanie, zásah, útvar
handler	zariadenie, jednotka, manažér
item	poznámka, prvok, bod, vec, jednotka, článok, odstavec, číslo, položka, člen, slovo, rovnako, detail, zápis, odsek
scaler	reduktor, čítač
timer	časovač, rozdeľovač, krokovač, hodiny, jednotka, čas
unit	jednotka, celok, oddiel, útvar, jednička, jedna, zväz, diel, agregát, stroj, prístroj, zariadenie, lístok, ročník, kurz, terminál, kus, skupina, zložka, blok, člen, spojenie
unity	zhoda, súdržnosť, zomknutosť, jednotka, jednota

Table 1. Forward and backward translation of word *jednotka*

#	Intersection based on Slovak word	English word whose translation contains specified Slovak word
1	celok	entity, unit
2	zariadenie	handler, drive, unit
3	člen	element, item, unit
4	útvar, oddiel	force, unit
5	prvok, článok	element, item

Table 2. Result of backward Slovak translation

#	Slovak synset	English synset
1	jednotka, celok	entity, unit
2	jednotka, zariadenie	handler, drive, unit
3	jednotka, člen	element, item, unit
4	jednotka, útvar, oddiel	force, unit
5	jednotka, prvok, článok	element, item

Table 3. Linking Slovak and English synsets

4.3 Using thesauri

Thesauri, as mentioned above, are natural source of synsets. We can use them similarly as we used WordNet database – taking the set of synonyms, translating each word and intersect particular translations. However, different approach can be used if we have thesauri for both languages. In this case we can try to find corresponding synsets and link them together.

4 Conclusion

At the moment of manuscript preparation we have implemented two approaches:

1. building synsets based on WordNet using translation dictionaries [6];
2. building synsets using on-line translation dictionaries only [7].

Experiments performed by implemented approaches showed that none of them is perfect. Each of them produced very good result for some words but at the same time very poor results for other words. Results can be affected by several factors:

- word used to specify synsets are usually not true synonyms; they often represent close senses only;
- translation dictionaries do not contain all synsets;

- different authors (editors) presents different subsets of word constituting synsets in their dictionaries;
- no dictionary specifies all senses of a word (for most of the words);
- two-way dictionaries (especially printed versions) are not symmetrical (we can not be sure that for translation $w^{L_1} \rightarrow w^{L_2}$ we will find translation $w^{L_2} \rightarrow w^{L_1}$

In the future we plan to integrate presented approaches in the common single approach which will be used for building relevant linked synsets.

References

[1] WordNet. A lexical database for the English language. http://wordnet.princeton.edu/

[2] EuroWordNet. A multilingual database with wordnets for several European languages. http://www.illc.uva.nl/EuroWordNet/

[3] Balkan WordNet (BALKANET). http://www.ist-world.org/ProjectDetails.aspx?ProjectId=a137b147dead4b75b11d4d8da46e7767

[4] Online Dictionary. Slovensko-maďarský slovník http://slovnik.agx.sk/dictionary_main_sk_hu.php

[5] Translation dictionary FEEHA. http://www.feeha.sk/translator.php

[6] Lapoš P.: Overenie budovania EuroWordNet synsetov na báze on-line slovníkov. Diploma thesis. KPI FEI TU Košice. 2005. In Slovak.

[7] Sudynová M.: Nástroj na generovanie slovníkových záznamov. Diploma thesis. KPI FEI TU v Košiciach. 2006. In Slovak.

Shallow Ontology Based on VerbaLex

Marek Grác

NLP Centre, Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

Abstract. Ontologies have proven to be a useful resource in natural language processing. In this paper, we introduce basic ideas of a shallow ontology named Sholva. This ontology is based on VerbaLex, a database of verb valencies, where each valency pointer also contains a pointer into EuroWordnet. We focused our effort on building ontology which would help us in solving real problems in syntactic analysis, word sense disambiguation and machine translation.

1 Introduction

The information around us is increasing with every second. One is no longer able to read and process it in a lifetime. Processing of natural language on computers can help us in finding the most relevant pieces. Various resources are used for this purpose. Semantic networks are among the most popular formalisms for knowledge representation. Like other networks, they consists of nodes and links. For English we have a number of possibilities from domain oriented to general ones like Princeton WordNet [8], for Czech one big general semantic network EuroWordNet [12]. It is very rich and complex network but unfortunately only few applications use its potential. Creation of similar resources for other close languages such as Slovak is very difficult and also time-consuming.

Our goal is to create a simpler ontology which will be easy to create and use primarily in our existing applications. This application-driven approach should help us to avoid creating a perfect complex ontology by providing us with a simple one instead, which can be used in various projects right now. Simpler ontology should also help us to create similar resources for other languages and take advantage of it in machine translation. Such project can reuse many existing components that were created for different purposes and projects.

2 WordNet ontologies

A lexicon with information about how words are used and what they mean is a necessary component for any application working with natural language. Ontologies are one of the resources that can provide enough information for those. Ontology is a formal representation of a set of concepts within a domain and the relationship between those concepts. Ontologies can be based on different assumptions, for specific domains and different purposes. Thus it is very difficult to compare them using objective metrics.

There are several ontologies built for the English language. For smaller European languages, one of the most important general ontologies is Princeton Wordnet [8]. It

contains many relations (e.g. hypo/hyperonym, is part of) connecting synsets (synonym set) which are equated with "senses". Specifically, according to WordNet's on-line glossary, a *sense* is a "a meaning of a word in WordNet. Each sense of a word is in a different synset". Princeton WordNet is available under free license also for commercial applications.

EurowordNet [12] and Balkanet [11] were projects to localize (and improve) parts of the original version to Central and South East European languages. Thanks to ILI (inter lingual index), it is possible to connect ontologies and use the result as a multilingual dictionary. Unfortunately some of the problems of original WordNet still remain [4], e.g. the assumption that membership in two or more synsets is equivalent to having more different senses. Some of the WordNet senses are indistinguishable from one another by any criterion. Attempt to build a WordNet-like ontology for new language was described in various papers [9], [1]. Creation of proper synsets and assigning the relations is a time-consuming process that needs expert in this field. One of the most serious problems of the EWN data is their very strict license.

3 VerbaLex

The basic sentence frame is driven by the lexical characteristics of its predicative constructions based on the set of possible verb valencies of the particular verb. For Czech languages we posses several lexicons of verb valencies. The first resource is Czech WordNet valency frames dictionary, which was created during Balkanet project and contains semantic roles and links to the Czech WordNet. The other resource, Vallex 1.0 [3] is a lexicon based on the formalism of the Functional Generative Description (FGD) and was developed during Prague Dependency Treebank [2] project. Latest project named VerbaLex [5] comprehends all the information found in the previous resources plus adds other relevant ones such as verb aspect, verb synonymy and semantic verb classes based on VerbNet project [10]. VerbaLex contains 10 478 verbs, 21 123 verb senses and 19 360 valency frames. Information in VerbaLex is stored in the form of *complex valency frames* (CVF).

Complex valency frame is designed as a sequence of elements which form a list of necessary grammatical features (e.g. preposition and grammatical case).

opustit:4/leave office:1 (give up or retire from a position)
frame: **AG** <*person:1*>$^{obl}_{who1}$ **VERB ACT**<*job:1*>$^{obl}_{what4}$
example: Jarek opustil zaměstnání / Jarek left his job

Example sentence can show us that if "Jarek" has to be the agent (semantic role) then it has to be in nominative (numbered 1) case. Also it has to be a hyponym of person:1 in the WordNet ontology. Thanks to ILI we can have nodes named in English and use words from Czech EuroWordNet.

This notation exported to an XML format allows us to easily process both syntactic and semantic layer of the sentence.

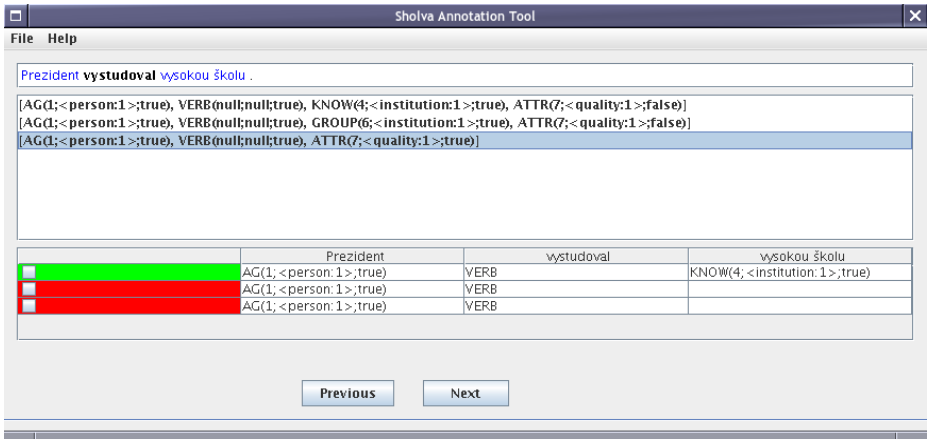
4 Syntactic analyzer synt

The NLP laboratory at FI MU Brno develops a deep syntactic analyzer of Czech sentences, `synt` [6]. The system uses a meta-grammar formalism which enables it to define a grammar with maintainable number of meta-rules. These meta-rules are produced manually by linguists. Meta-rules are translated into context-free rules supplemented with additional contextual constraints and semantic actions. Efficient and fast head-driven chart parsing algorithm is used for context-free parsing. Final result of parsing is a set of derivation trees – unfortunately, for more complicated sentences we can obtain billions of those. After modification of the parser and grammar [7] it is possible to obtain syntactic structures corresponding to the given nonterminal. With this feature it is possible to use this tool not only as a deep but also as a shallow parser which identifies most of important syntactic structures unambiguously (as opposed to the possibly huge number of derivation trees).

5 Sholva

The Sholva ontology, currently being developed at FI MU, attempts to create a new lexical resource for Czech language which will be free to use for any purpose. Proposed methods and implementation of the tools are also suitable for other languages with similar structure (e.g. Slovak) and we believe that those languages will follow soon. Our ontology should be used primarily for machine translation, syntactic parsing and word sense disambiguation. Application driven extensions should be possible.

Our attempt is based on an assumption that creation of a rich lexical resource is time-consuming and very often the subjective view is one of the most serious problems. Thus our methodology aims to limit the expert (linguist) as much as possible. We believe that language constructions are obtainable from corpora, so only usages found in corpora are taken into account.



We do not intend to create an ontology with dozens of relations and complicated structures and we would rather like to follow the KISS methodology (Keep It Simple, Stupid). For our usage, only hypo/hyperonymical relations can be used directly, and basic ontology will not contain any other relations. In EWN, senses of words are numbered, but splitting word senses is also a very subjective task. More importantly, this has no direct relevance to our primary goals. For word sense disambiguation, we need to be able to distinguish various senses of a given word. We believe that for this purpose, the knowledge of path from the root node to any given word created by hyponymical relations defines *word sense*. It is possible that a word will have several hyperonymical relations, but we do not know if they refer to the same or a different thing.

We know where we will obtain data for processing and which relations we will focus on. The process of creating an ontology in this style is very similar to corpora annotation. Annotation tool will provide everything to the expert who will just select the correct option or mark an error (syntactic analyzer error or a missing valency frame). Users will have to select line(s) which contain correct valency and mapping of verb roles to syntactic structures (mostly preposition and noun phrases). Lines which fill all obligatory slots are marked green, all others are marked red (and cannot be selected). Problems may occur if a slot contains logical OR (e.g. abstraction:1 or communication:2 or info:1), in which case, we will just assign this formula to the syntactic structure. Later, when our structure will occur in a more specific slot, we can disambiguate it. It may also be possible that the given word does not need to be distinguished at all and is only assigned to a slot within this formula. In that case, the ontology hierarchy and valency frames should be modified to better correspond to the language usage.

6 Conclusions

In this article about Sholva, Shallow Ontology Based on Valency Frames, we have presented the basic ideas. Moreover, motivation and reasons for different ontology style have been explained and a methodology for building such an ontology on top of existing resources has been given.

In the future there will be further work on annotating large enough corpora – this should have positive effects on quality of VerbaLex valencies and parsing of syntactic structures in `synt`. Existence of such lexical resource suitable for machine processing should help in our major areas of interest – syntactic parsing and machine translation. Moreover, after creation of a similar resource for other languages, it will also be possible to create a multilingual dictionary.

7 Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the project 1ET100300419 and by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009.

References

[1] Erjavec, T. and Fišer, D. (2006). Building Slovene WordNet. In *5th Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

[2] Hajič, J. (1998). Building a syntactically annotated corpus: The Prague dependency treebank. *Issues of valency and meaning*, pages 106–132.

[3] Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolár, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Largecoverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, pages 57–68. Citeseer.

[4] Hanks, P. and Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue française de linguistique appliquée*, (2005/2):63–82.

[5] Hlaváčková, D. and Horák, A. (2005). Verbalex–new comprehensive lexicon of verb valencies for czech. In *Proceedings of the Slovko Conference*. Citeseer.

[6] Horák, A. and Kadlec, V. (2006). Platform for Full-Syntax Grammar Development Using Meta-grammar Constructs. In *Proceedings of the Paclic*.

[7] Jakubíček, M., Horák, A., and Kovář, V. (2009). Mining Phrases from Syntactic Analysis. In *Text, Speech and Dialogue*.

[8] Miller, G. (1995). WordNet: a lexical database for English.

[9] Pala, K. and Smrž, P. (2004). Building Czech wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2):79–88.

[10] Schuler, K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. *Univ. of Pennsylvania-Electronic Dissertations*.

[11] Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiș, D., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2002). BALKANET A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25.

[12] Vossen, P. (1998). Eurowordnet a multilingual database with lexical semantic networks. *Computational Linguistics*, 25(4).

Multimodal Russian Corpus (MURCO): General Structure and User Interface

Elena Grishina

Institute of Russian Language RAS, Moscow, Russia
rudi2007@yandex.ru

Abstract. In the paper the Multimodal Russian Corpus (MURCO) is described in outline. The MURCO is planned to become accessible for general use at the beginning of 2010. Just now the group of the Corpus developers prepares the MURCO pilot project. Therefore, the discussion concerning its general structure and interface characteristics is quite urgent.

1 What is MURCO?

The MURCO is being elaborated within the framework of the Russian National Corpus (RNC) and may be considered as the further development of the spoken component of the RNC. The latter includes the Spoken Subcorpus proper and the prosaic part of the Accentological Subcorpus[1]. The Spoken Subcorpus is the collections of transcripts of the spoken texts of different types (private speech, public speech, and movie speech-tracks) (see [1], [2], [5], [12]). These transcripts are annotated morphologically and semantically according to the RNC annotation system (see [14], [18] [16]), and also they are annotated from the metatextual point of view (see [17]). In addition, the Spoken Subcorpus has its own annotation: the accentological annotation and the sociological one. The ***accentological annotation*** presupposes that in every wordform the real (not normative) stress is marked. Therefore, a user can investigate the history of Russian accentological system and its normative requirements, which are specific for this or that period. The ***sociological annotation*** means that to every token the information concerning the sex, the age and the name of a speaker is assigned, so a user can form his own subcorpora according to all these parameters and their combinations.

Thus, the Spoken Subcorpus of the RNC (its volume just now is circa 8,5 million tokens) gives a user the possibility to solve different types of tasks (lexical, morphological, semantic, accentological, sociological and mixed), but all these tasks must not be connected or based on the real phonation. Therefore, the problems concerning phonetics, orthoepy, intonation, and also the morphological, semantic, accentological, sociological problems, which are founded on any aspect of phonetics, have no chance to be solved by means of the data of the Spoken Subcorpus of the RNC.

[1] The poetic part of the Accentological Subcorpus includes Russian poetry of 18–21 centuries with the marked arses (potentially stressed syllables). It gives a user the possibility to figure out the real stress of a wordform according to the simple set of the rules (see [6], [10]).

Such are the reasons for the creation of the MURCO. The main principle of the Multimodal Russian Corpus is the alignment of the text transcripts and the parallel sound- and video tracks (see [11], [7], [8]). Consequently, when a user makes his data query he may obtain not only written text, annotated from different points of view, but also the corresponding sound- and video material. This possibility let an investigator use the obtained information at his will: he may make use of his own manner of phonetic transcription, of the elected speech and intonation analyzers; he may pose and solve all types of research tasks, connected with phonetics, and so on.

Quite a conundrum for the developers of the multimodal corpora is the whole set of problems concerning copyright offence and privacy invasion (see proceedings of [15]: every paper, pertinent to the multimodal subject area, contains the debate on the point). It is the main reason why the multimodal corpora usually are not accessible for general use. To avoid the range of the copyright and privacy problems, we have decided to develop the MURCO on the basis of movie subcorpus of the Spoken Subcorpus of the RNC[2]. There are some features, which distinguish the natural spoken speech and the cinematographic one (first of all we mean the parameter of the text coherence), but the differences though remarkable are not crucial (see [3], [4] about the usage of the discourse markers in the movie transcripts; the strategy of their usage is virtually the same in the natural and cinematographic spoken Russian); that is to say, the higher coherence of the movie transcripts in comparison with the transcripts of the natural spoken texts does not turn the former into the written texts: they remain spoken ones.

The total volume of the movie transcripts in the RNC is circa 5 million tokens. We plan to include the movie subcorpus in the MURCO wholly, without reservation. As a result we mean to obtain the first-rate generally accessible multimodal corpus.

2 MURCO annotation

The types of annotation in the MURCO have been described in our earlier papers (see [11], [7], [9]), so in this paper we only summarize the theme. Let alone the standard RNC annotation (see above), the MURCO annotation includes the following items:

- orthoepic annotation: combinations of sounds are marked
- annotation of accentological structure: the word structure from the point of view of stress position is defined
- speech act annotation: the types of speech acts and vocal gestures, used in a clip are described
- gesture annotation: the type of gesticulation in a clip is described

[2] Later on we plan to include in the MURCO the patterns of the public spoken speech, which also does not involve the problems of the kind.

So, the summarized scheme of the MURCO annotation seems to be as follows:

Method of annotation *Assigned*	Automatic (obligatory)	Semi-automatic (obligatory)	Manual (selected)
To text	metatextual annotation	–	–
To word	morphological, semantic, orthoepic annotation, annotation of accentological word structure	sociological, accentological annotation	–
To clixt (text+clip)	metatextual annotation	–	speech act and gesture annotation

Table 1.

Naturally, the types of annotation, which are specific for the MURCO (in the Table 1 they are spaced out), define the peculiarities of the MURCO interface in comparison with the interface of the RNC proper. So below we'll describe only the items of the MURCO interface, which differ from the RNC interface.

3 Characteristics of MURCO interface

3.1 Orthoepic queries

The orthoepic annotation in the MURCO is founded on the morpho-phonemic principle of the Russian orthography, which means that there are quite transparent correspondence between the word and word-combination spelling and the word pronunciation. Therefore, we receive the possibility to annotate the combinations of letters to obtain the pronunciation of the correspondent sounds.

The crucial types of sound combinations in Russian are as follows:

- C…C = combination of two or more consonants within the word limits
- V…V = combination of two or more vowels within the word limits
- C…C#C…C = combination of consonants at the word boundaries
- V…V#V…V = combination of vowels at the word boundaries
- C…C#V…V = combination of the consonants before the vowels at the word boundaries

Obviously, it is quite easy to annotate such combinations of letters in a text automatically. Consequently, to any tokens in the MURCO will be assigned the set of the letter combinations, which may present a difficulty. Naturally, all these combinations suppose to become searchable.

For example, we may formulate our query something like this: *NS* (that is, 'to find all tokens, which include the letter combination -*ns*-), and receive as a result the clixts, in which may be found tokens like this: *pensija* 'pension', *pansion* 'boardind-house', *Pensil'vanija* 'Pennsylvania', *mansarda* 'attic floor', *tangens* 'tangent' (and their derivatives and indirect forms), and so on. If we request *N#S* (that is, 'to find all tokens, which have the letter combination -*n*- and -*s*- at their boundaries'), we receive clixts with the combinations of tokens like this: *on sejchas* 'he now', *voron sel* 'a raven has sat', and so on. Obviously, it is very hard to obtain the tokens including the letter combinations of the kind without special orthoepic annotation. From the other hand, the queries of the kind may be very useful. For example, the query described above gives us the possibility to investigate the pronunciation of the letter combination -*ns*- before the front vowels. It is well known that in Russian its pronunciation may be of two kinds: [n's'] with palatalized [n'] and [ns'] with hard [n]. The orthoepic annotation assists us to define the main factors, which have an influence on the choice between these two possibilities (the choice depends on the degree of the assimilation of this or that word, on the age of a speaker and on the presence/absence of the word boundary between two consonants).

3.2 Queries on accentological word structure

It is well known that the dynamic quality of Russian stress leads to the great degree of the reduction of the unstressed syllables in a word. Consequently, it is very important to give a user an opportunity to obtain information, concerning the position of the stressed syllable and the quality of the stressed vowel, the position/quality of the pre- and post-tonic vowels, and so on. Owing to the fact that the majority of the clixts in the MURCO are accentuated, it is possible to annotate the accentological structure of any token in automatic mode. The content of the possible requests is defined in line with the Table 2:

	quality of vowel	number of syllable
stressed vowel	A	1
pre-tonic vowel	B	2
post-tonic vowel	C	3
	quantity	
syllables	4	

Table 2.

In the table cells A–C a user may specify the letter designation of a vowel (in the stressed, pre- and post-tonic syllable), in the cells 1–3 – the number of the corresponding syllable, in the cell 4 – the quantity of the syllables in a word. All these parameters are independent, so a user can freely combine them if necessary. For example, a user may request all tokens containing 1) the second post-tonic syllable, 2) the stressed syllable *o*, 3) three syllables, 4) vowel *o* in the second pre-tonic syllable, while a token has 4 syllables and the stressed vowel *o*.

All these parameters are very important for the phoneticians, specialists in orthoepy, dialectologists, and investigators in the area of the history of Russian. In addition, the importance of orthoepic and accentological annotation can scarcely be overestimated, having in mind the professional interests of the teachers of Russian, uppermost as a foreign language.

3.3 Speech act queries

The specialists on Russian intonation, linguistic pragmatics, socio- and psycholinguistics will obtain using the MURCO very significant information on different aspects of the speech acts. The acquisition of the information of the kind is provided with the special annotation, which is attached to a clixt in manual mode. Earlier we have described the basic principles of this annotation (see [7], [8]) and the software to hasten the process of the annotation and to unify its results (see [13], [9]).

In this paper we would like to present the items of the MURCO interface, connected with this type of annotation.

3.3.1 Sociolinguistic characteristics of clixt

- Quantity of participants (1, 2, 3, many)
- Participants' sex (Mas, Fem, Mixed)
- Language (Russian, Russian with accent, Foreign (Ukrainian, English, and so on), Quasi, Secret… the list is open)
- Social situation (Telephone call, Dinner speech, Talk with authorities… the list is open)

3.3.2 Types of speech acts in clixt

1. Address or call
2. Agreement
3. Assertion
4. Citation
5. Complimentary
6. Critical utterance
7. Etiquette formula
8. Imperative
9. Joke
10. Modal utterance or performative
11. Negation
12. Question
13. Trade utterance

Naturally, every type has the set of its own subtypes, for example,

- ***Question***
  - Closed
  - Open
  - Inarticulate
  - Overinterrogation
  - Feedback
  - Question to oneself
  - Leading
  - Indirect
  - Contact
  - Incomprehension
  - Critical
  - Etiquette
- ***Joke***
  - Irony
  - Hint
  - Sneer
  - Joke

Till the moment we pick out 150 subtypes of different speech acts grouped into 13 types, mentioned above. The majority of the speech act types are based on the Russian verbs, describing the process of speech. Consequently, they are intuitively comprehensible by all Russian speakers. To every clixt may be attached more than one type of speech act, and moreover, every speech act in a clixt may be characterized from different points of view (i.e., an assertion may be characterized at the same time as information, declaration, statement, and some others). Thus, the classification of speech act is not tree-like, it is faceted.

3.3.3 Completeness of utterances

- Full
- Self-interruption
- Question without answer
- Overlapping cues
- Unfinished
- Interrupted
- Continued
- Gesture instead of word

3.3.4 Repetitions and their types

- No repetitions
- Types of repetitions
  - Single
  - Multiple
  - One-word
  - Many-word
  - Overinterrogation
  - Repetition with intensifier
  - Repetition with different intonation
  - Echo
  - Relay
  - Mimicking
  - Repetition with the change of listener
  - Envelope repetition

3.3.5 Manner of speech

- Normal
- All together
- Articulation disorders
- Chanting/scanning
- Crying
- Declamation
- Dictation
- Drunken
- Dubbing-in
- Fast
- Humming
- Inarticulate
- Laughing
- Parceling out
- Reading
- Shout
- Slip of the tongue
- Talking to oneself
- Ventriloquism
- Voice-over
- Whisper

3.3.6 Vocal gestures and interjections

- Vocal gestures
  - Address to an animal
  - Beep
  - Click of the tongue
  - Feeling cold
  - Hemming
  - Iconic sounds
  - Stopgaps
  - Teasing sounds
- Interjections
  - Admiration
  - Agitation
  - Agreement
  - Approval
  - Attraction of smb's attention
  - Backing-hum
  - Backing-well
  - Backing-yes
  - Backing-вот
  - Bewilderment
  - Chagrin
  - Comprehension
  - Dissatisfaction
  - Distrust
  - Fright
  - Grief
  - Indignation
  - Mockery
  - Negation
  - Pain
  - Pity
  - Question
  - Recollection
  - Reply to address
  - Scorn
  - Sudden recollection
  - Surprise
  - Threat
  - Unexpectedness
  - Whoops!
- Physiological activities
  - Breathe the air
  - Chuckle
  - Cough
  - Exercise stress
  - Exhalation
  - Groaning
  - Kiss
  - Sigh
  - Smacking one's lips
  - Sniffing
  - Spit
  - Taking one's breath
  - Weeping
  - Whistle
  - Yawn

These are the main point of the MURCO interface, concerning the speech act annotation. As follows from the written above, the MURCO let a user collect very diverse material as for the speech activities. This fact opens up new possibilities for the investigators of spoken Russian.

3.4 Gesture queries

During last three decades the investigation of the role of the gesticulation in different languages has progressed to a large degree. Now it is the current opinion that it is time to elaborate the gesture corpora to base the investigation of the gesture systems on a hard ground (see the materials of [15] and their review and the main bibliography in [7]).

The MURCO seems to be the resource, which is generally accessible and quite considerable as for its volume, moreover, the MURCO is planned to include a lot of video tracks. So, it is absolutely necessary to provide a user with the annotation and interface concerning Russian gesticulation.

The basic principles and ideological grounds for our gesture classification we gave described earlier (see [7]). So, in this paper we list the main items of the MURCO interface, concerning the gesticulation subject matter.

3.4.1 Sociological characteristics of gesture

- Name of a speaker (if known)
- Speaker's sex (Mas, Fem, Unknown)
- Character's sex (Mas, Fem, Unknown, Man, playing woman, Man, pretending to be woman, Woman, playing man, Woman, pretending to be man)
- Speaker's age (baby, teenager, adult, elderly, unknown)
- Character's age (baby, teenager, adult, elderly, unknown)

3.4.2 Repetition factor

- Single gesture
- Multiple gesture

3.4.3 Active organ

- Main organ: head
  - brow
  - brows
  - chin
  - ear
  - eye
  - eyes
  - face
  - forehead
  - head
  - lips
  - lower lip
  - mouth
  - nose
  - tongue
  - upper lip
  - upper teeth

- Main organ: body
  - body
  - shoulder
  - shoulders
  - back
- Main organ: arm
  - arm
  - fingers
  - forefinger
  - forefinger+long finger
  - forefinger+long finger+fourth finger
  - forefinger+long finger+thumb
  - forefinger+thumb
  - fourth finger
  - hand
  - little finger
  - long finger
  - thumb
- Main organ: arms
  - arms
  - hands
  - forefingers
  - fingers
- Main organ: leg
  - foot
  - shin
- Main organ: legs
  - feet
  - legs

3.4.4 Passive organ

The set of the passive organs is specific for this or that active organ. The basic passive organs are as follows:

- No passive organ
- arm
- arms
- back
- body
- breast/stomach
- chin
- eat
- eye
- face
- fingers
- hair
- hand
- head
- hip
- hips
- lips
- lower lip
- mouth
- neck
- nose
- shoulder
- throat

3.4.5 Adaptor

Adaptor is the object, which is a necessary component of this or that gesture, but is not one of the human organs. The main types of adaptors are as follows:

- No adaptor
- cloth
- earth
- external object
- glasses
- gloves

- handset
- headwear
- heavy object
- interlocutor
- piece of furniture
- pocket
- sky
- surface
- tableware
- tie
- vessels
- watch
- wristlet

3.4.6 Dimensional characteristics of gesture

- Palm orientation
  - up
  - down
  - one opposite the other
  - to speaker's body
  - outside
  - perpendicularly to speaker's body
- Direction of movement
  - backward
  - differently directed
  - does not matter
  - downwards
  - forward
  - forward-backward
  - from right to left
  - from the outside to the center
  - from within outside
  - horizontal circle
  - on its axis
  - outside
  - to oneself
  - to the center
  - upwards
  - vertical circle

3.4.7 Gesture meanings and gesture types

Till the moment we have marked out about 250 gesture meanings, which are grouped into 14 gesture types. The gesture types are as follows (see at greater length in [8]):

- Adopted
- Conventional
- Corporate
- Critical
- Decorative
- Deictic
- Etiquette
- Gestures – speech acts
- Gestures of inner state
- Iconic
- Physiological
- Regulating
- Rhetorical
- Searching

Every type includes some gesture meaning. For example, some of the etiquette gestures are as follows:

- gratitude (*to applaud, to move one's head forward, twice-repeated kiss, to close one's eyes, to nod, to touch smb, to bow, to touch smb's hand, to kiss smb, to kiss smb's hand, press one's hands to one's breast*, and so on)

- apology (*to beat one's breast, to nod, to move one's chin outside, to press smb's hand to one's breast, to press one's hand to one's breast*)
- invitation (*to nod, to show smth with one's hand, to bow*), and so on.

4 Conclusions

Thus we can see that the MURCO considerably extends searching possibilities up about the characteristics of spoken Russian. We may illustrate the fact with the queries, applying to the Russian greeting formulas (GF).

Corpus Types of queries	**Spoken Subcorpus of RNC**	MURCO
(3) **Lexical queries**: the retrieve of the specific lexemes, used in GF (i.e. *zdravstvujte* 'how do you do?', *privet* 'hi!', and so on)	+	+
(4) **Morphological queries**: the retrieve of the specific morphological characteristics of the GF lexemes (i.e. *zdravstvujte* (Pl or courtesy) vs *zdravstvuj* (Sg), *privet* (Noun) vs *privetstvuju* (Verb), and so on)	+	+
(5) **Sociological queries:** the forming of the gender and chronological subcorpora to investigate the peculiarities of the GP usage	+	+
(6) **Semantic & speech act queries:** the retrieve of all Russian GP simultaneously	–	+
(7) **Orthoepic/accentological queries:** the retrieve of the types of the vowel contractions and the shortening of the consonant groups in the GF; the reduction of the pre- and post-tonic vowels in GF	–	+
(8) **Speech act queries:** the retrieve of the types of repetitions, used in GF; the types of vocal gestures and interjections, accompanying the different types of GF; GF, used in the man/woman dialogues; see also the item 4	–	+
(9) **Gesture queries:** the retrieve of the gestures, accompanying Russian GF	–	+

Table 3.

Acknowledgement

The work of the MURCO group is supported by the program "Genesis and Interaction of Social, Cultural and Language Communities" of the Russian Academy of Sciences. The author's investigation is supported by the RFBR[3] (RFFI) under the grant 08-06-00371a.

References

[1] Grishina, E. (2005). Ustnaja rech v Nacional'nom korpuse russkogo jazyka. *Nacional'nyj korpus russkogo jazyka: Rezul'taty i perspektivy.* M.: 94–110, available at: http://docs.google.com/Doc?id=df52fjjj_12fcbpqbdq&hl=en

[2] Grishina, E. (2006). Spoken Russian in the Russian National Corpus (RNC). *LREC2006: 5th International Conference on Language Resources and Evaluation*. ELRA: 121–124, available at: http://docs.google.com/Doc?id=df52fjjj_3wd9mcrdg&hl=en

[3] Grishina, E. (2007). O markerah razgovornoj rechi (predvaritel'noje issledovanije podkorpusa kino v Nacional'nom korpuse russkogo jazyka). *Kompjuternaja lingvistika i intellektual'nyje tehnologii. Trudy mezhdunarodnoj konferencii "Dialog 2007".* M.: 147-156, available at: http://docs.google.com/Doc?docid=df52fjjj_0d5gbwscn&hl=en, http://www.dialog-21.ru/dialog2007/materials/html/22.htm

[4] Grishina, E. (2007). Text Navigators in Spoken Russian. *Proceedings of the workshop "Representation of Semantic Structure of Spoken Speech" (CAEPIA'2007, Spain, 2007, 12-16.11.07, Salamanca)*. Salamanca: 39-50, available at: http://docs.google.com/Doc?docid=df52fjjj_11fmxsdzdh&hl=en

[5] Grishina, E. (2008). Ustnaja rech v Nacional'nom korpuse russkogo jazyka. *Sbornik trudov. XX sessija Rossijskogo akusticheskogo obsh'estva.* M.: 88–91, available at: http://docs.google.com/Doc?docid=df52fjjj_2gf3httgq&hl=en

[6] Grishina, E. (2009). Korpus "Istorija russkogo udarenija". *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novyje rezul'taty i perspektivy.* M.-SPb.: 150–213

[7] Grishina, E. (2009). Mul'timedijnyj Russkij Korpus (MURCO): problemy annotacii. *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novyje rezul'taty i perspektivy.* M.-SPb. (forthcoming)

[8] Grishina, E. (2009). Nacional'nyj korpus russkogo jazyka kak istochnik svedenij o russkoj rechi. *Rechevyje tehnologii*. (3). M. (forthcoming)

[9] Grishina, E. Multimodal Russian Corpus (MURCO): types of annotation and annotator's workbenches. *Proceedings of Corpus Linguistic Conference 2009 (UK, Liverpool, 20–23 July)*. Liverpool (forthcoming)

[3] The Russian Fund of Basic Researches.

[10] Grishina, E., Korchagin, K., Plungian, V., Sichinava, D. (2009). Poeticheskij korpus v ramkah Nacional'nogo korpusa russkogo jazyka: obsh'aja struktura i perspektivy ispol'zovanija. *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novyje rezul'taty i perspektivy.* M.-SPb.: 71-113

[11] Grishina, E, Savchuk, S. (2008). Korpus zvuchash'ej russkoj rechi v sostave Nacional'nogo korpusa russkogo jazyka. Proekt. *Kompjuternaja lingvistika i intellektual'nyje tehnologii. Trudy mezhdunarodnoj konferencii "Dialog 2008".* M.: 125-132, available at:
http://docs.google.com/Doc?docid=df52fjjj_1dzj8gchs&hl=en,
http://www.dialog-21.ru/dialog2008/materials/html/19.htm

[12] Grishina, E., Savchuk, S. (2009). Ustnyj korpus v Nacional'nom korpuse russkogo jazyka: sostav i struktura. *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novyje rezul'taty i perspektivy.* M.-SPb (forthcoming)

[13] Kudinov, M., Grishina, E. (2009). Instrumenty poluavtomaticheskoj razmetki dl'a Mul'timedijnogo russkogo korpusa (MURCO). *Kompjuternaja lingvistika i intellektual'nyje tehnologii. Trudy mezhdunarodnoj konferencii "Dialog 2009".* M.: 248–261, available at:
http://www.dialog-21.ru/dialog2009/materials/html/40.htm

[14] Kustova, G., Lyashevskaja, O., Paducheva, E., Rahilina, E. (2005). Semanticheskaja razmetka leksiki v Nacional'nom korpuse russkogo jazyka. *Nacional'nyj korpus russkogo jazyka: Rezul'taty i perspektivy.* M.: 155–174, available at:
http://ruscorpora.ru/sbornik2005/10kustova.pdf

[15] LREC (2008). *6th International Conference on Language Resources and Evaluation*. Marrakesh: ELRA. Available at:
http://www.lrec-conf.org/proceedings/lrec2008/

[16] Lyashevskaja, O., Plungian, V., Sichinva, D. (2005) O morfologicheskom standarte Nacional'nogo korpusa russkogo jazyka. *Nacional'nyj korpus russkogo jazyka: Rezul'taty i perspektivy.* M.: 111-135, available at:
http://ruscorpora.ru/sbornik2005/08lashevs.pdf

[17] Savchuk, S. Metatekstovaja razmetka v Nacional'nom korpuse russkogo jazyka: bazovyje principy i osnovnyje funkcii. *Nacional'nyj korpus russkogo jazyka: Rezul'taty i perspektivy.* M.: 62-88, available at:
http://ruscorpora.ru/sbornik2005/05savchuk.pdf

[18] Sichinava, D. (2005). Obrabotka tekstov s grammaticheskoj razmetkoj: instrukcija razmetchika. *Nacional'nyj korpus russkogo jazyka: Rezul'taty i perspektivy.* M.: 136-154, available at:
http://ruscorpora.ru/sbornik2005/09sitch.pdf

Electronic Lexical Card Index for the Ukrainian Dialects (ELCIUD)

Pavlo Grytsenko, Olena Siruk, and Viktor M. Sorokin

Institute for the Ukrainian Language, Ukrainian Academy of Sciences
Laboratory for Computational Linguistics,
National Taras Shevchenko University of Kyiv, Ukraine
olebosi@gmail.com

Abstract. The aim of this paper is to present an Electronic Lexical Card Index for the Ukrainian Dialects (ELCIUD). Our investigation focuses on the main problems that occur during compilation process (searching of the optimal structure of a database and an electronic dialect card, granting a possibility of editing all the linguistic information inside and adding a new one taking into account differences in localization marking systems of different authors etc.) and possible ways of resolving them. We also review future trends of ELCIUD development, especially its transformation into a computational lexicographical system, and using it as a base for further linguistic experiments as computer means of dialect text research.
Keywords: dialect, lexical databases, computational lexicographical systems.

1 Topicality of the research

The computer-aided design of lexicographic systems satisfying the demands of today's information-aware society for the systematized and freely accessible linguistic data is an important task of modern Ukrainian lexicography. The goal is to get comprehensive and detailed information quickly into a form suitable for its processing in accordance with the formulated linguistic task. A computer dictionary, providing at least the same degree of thoroughness of a conventional paper dictionary, enables us to extend substantially the information scope of the lexicographic system, to better prepare language material for subsequent processing and in general to increase the speed of work. Consequently, the design of computer dictionaries, card indexes and additional software is of pressing concern.

It is a question of great importance not only for literary idiom, but also for dialects. In recent years dialectology has taken up the use of electronic sources of information (such as electronic dialectal atlases, text corpora of interviews with respondents, postal questionnaires and historical text corpora) and electronic methods (such as computer mapping, statistical processing of data, collation with atlases and presentation on maps)[1]. Thus attention is concentrated on **dialectometry**, defined as the mathematically

[1] The first attempts at the application of computer methods for language data processing were made by Slovak researchers, in particular P. Žigo for the "Slavic Linguistic Atlas". In the 1990s Žigo promoted the advantages of working with computers and

expressed degree of identity∼difference between the compared dialectal systems (with the employment of frequency analysis and orientation for psychological conclusions)[2] or on the closeness∼remoteness of the phenomena in compared dialects or dialectal languages [8]. It is self-evident that the application of mathematical factors during the comprehension of dialectal materials and the use of computer technologies brings with it increased perspectiveness and expedience. Similarly, the use of computer technologies for database preparation, modeling, and development of dialectal dictionaries as information generators with respect to a dialectal language brings great benefits [1]. The experience of computational lexicography [14], based largely on processing of the the literary or standard language, can also be of great help in state-of-the-art dialectal lexicography.

2 Theoretical principles of the project

Electronic Lexical Card Index for the Ukrainian Dialects (ELCIUD) is a linguistic database (LDB), created on the basis of published dialect dictionaries and unpublished

detailed his experience of collecting dialectal material in Slovakia [2]. Polish scholars began to use statistical methods for dialectological research at around the same time [10; 15]. Today in Poland there is an online information retrieval system dealing with dialects and dialectology: the project "Gwary polskie: Przewodnik multimedialny", edited by H. Karaś, was completed in 2009 [9]. The informatization of linguistic research is developing actively in the Russian Federation. In particular, the project "Lexical Atlas of Russian National Dialects" is characterized by a research methodology involving computer mapping and a systematic approach to the interpretation and cartographic presentation of the material [5; 6]. At the beginning of the 1990s, dialectal texts were seen as one of the basic resources of dialectology: dialects are investigated as separate communicative systems. This direction was referred to as "communicative dialectology". The concept of dialectal corpus creation as an optimal base for the maintainance and subsequent processing of dialect text material was utilised, as a dialectal subcorpus, in the composition of the National Russian Corpus of the Institute for the Russian Language (Russian Academy of Sciences) and the Saratov Dialectal Corpus (Saratov State University) [3; 4]. The understanding of dialectology in accordance with this approach can be traced all the way back to Aristotle, and, with the application of exact calculation methods and the latest technologies, has reached its highest development in the USA [11; 12; 13].

[2] The mathematical modelling of the areal variation of dialectal features in combination with a revision of the traditional understanding of a dialect, according to William Kretzschmar ("Quantitative Areal Analysis of Dialect Features"), has all the chances of updating our knowledge about language and its regional differences. The article is devoted to the quantitative analysis of separate dialectal features from the "Linguistic Atlas of the Middle and South Atlantic States". The XII International conference Methods XII (August 2005, Université de Moncton, Canada) was devoted to the problems of methods in dialectology, to innovative approaches and, in particular, contemporary theoretical and methodological trends in this area [16]. Within the conference special seminars were held by David Heap (University of Western Ontario), John Nerbonne (University of Groningen), William Kretzschmar (University of Georgia).

handwritten lexical card indexes of other projects (Ukrainian materials in the lexical part of the "Slavic Linguistic Atlas", the "Atlas of Ukrainian" and the "Lexical Atlas of Ukrainian").

The **goal** of the ELCIUD is to gather from different sources all lexical units of the Ukrainian dialects, to process these as elements of the information system and to provide operative access to this source of linguistic data. The task of the project carried out by the Institute for the Ukrainian Language of the Ukrainian Academy of Sciences and the National Taras Shevchenko University of Kyiv is the creation of the linguistic and computational complex ELCIUD, which allows **formalization** of the language material, and presentation of information (dialectal units) in a structured output, suitable for computer-aided processing.

3 Stages of work

The realization of the project is divided into three stages. In the **first** stage, the creation of a linguistic database of register units is to be made on the basis of dictionary material (included in the register are dialectal units of all parts of speech). This is the stage of conversion of present electronic versions of dictionaries into a database which includes: 1) title word; 2) the content part of the vocabulary entry. In the **second** phase we enlarge the database by the import of new dictionaries and separate dialectal units, watching out for and correcting conversion errors, and also correcting the text of vocabulary entries, without changing their content. At the **third** stage, formulation of the database structure is planned by the selection of different types of linguistic information. This is the stage of formulation of the explanatory part of vocabulary entries and their enrichment by additional linguistic information.

4 Issues with ELCIUD design

There are several issues which arise with the creation of ELCIUD and whose solution is of pressing importance. These can be presented in the following key questions:

1. what kind of linguistic information should be present in the database
2. how to present this information
3. how to add new information
4. what are the possible directions of further data processing
5. what format should the final product have.

Within regard to the **first** question, we decided to focus on the vocabulary entries of published dictionaries. However, this does not eliminate the possibility of also adding vocabulary entries from non-printed sources. The **second** problem required additional experimental research. Working with dialectal dictionaries revealed divergences in structure, filling and presentation of vocabulary entries. This occurred not only with different authors, but also for the same author [Arkushyn, Grytsak, Magrytska, Onyshkevych, SUSG]. For example, the structural elements of a vocabulary entry for a monosemantic noun in the dictionary of Polissya's dialects [Arkushyn] have clear enough formal signs and sequence of description (see Table 1).

№	**Structural element**	**Formal signs**	**Example 1**	**Example 2**
1	Headword	Word. Comes after an indentation. Given in capital letters.	ОБЛІПЛЯНЕЦЬ	ОБИТАЄМЦИ
2	Phonetic transcription	A word written in special symbols is in square brackets.	[обл'íпл'анец']	[обита́йемци]
3	Gen. sing. form (pl. – for Pluralia tantum)	Partial word. Begins with a hyphen, ends with a comma.	-нц'а,	-ив,
4	Marker of gender for a noun or a number (for Pluralia tantum)	Italicized lowercase letter(s); ends with a point.	*ч.*	*мн.*
5	Stylistic marker	Italicized lowercase letter(s); ends with a point.	*кулін.*	**absence**
6	Definition of lexeme	Sentence; begins with the character ‘, ends with the character ’ and has a point after it.	пиріжок з тертої картоплі	звірі, які постійно живуть на одному місці
7	Example of use	Text, written in a standard font, not italic, not bold, with the additional signs of phonetic transcription.	**absence**	То на́ши обита́йемци, живу́т в на́шому л'іси
8	Localization of dialecticism	Number from 1 to 601, in accordance with the list of localities produced by the author, or all of the populated localities.	463	244

Table 1. Correspondence between structural elements and their formal signs in a vocabulary entry for a monosemantic noun in [Arkushyn]

As we can see, some cells are irregularly filled (the text-example of the use of dialecticism is absent in the first example, also as a stylistic marker in the second example), and this irregularity is pervasive. When it concerns non-optional elements (e.g. stylistic marker), and basic elements (such as an example of the use of lexeme), and also when we have extremely syncretic representation of data in a vocabulary entry, it is not easy both to expand such article to its complete structure and automatize this process. The structure of the article within the limits of the similar part of speech can vary substantially, in particular depending on the number of values and pronunciations of the title lexeme. For example, vocabulary entries for the headword “baba”:

БАБА[1] [ба́ба] -и, ж. *1.* ‘жінка, яка має внуків’ (всі н. пп.); *2.* ‘будь-яка стара жінка’ (всі н. пп.); *3.* ‘дружина’. А твоя́ ба́ба вдо́ма? Вона́ ни бойи́ць’:а тако́го чолоўи́ка, йак ти́ 392, 11, 90, 342; *4. знев.* ‘дівчина’. К’іровн’ічка хо́че нас в тре́т’ому посади́ти з баба́ми. 392.
БАБА[2] [ба́ба] -и, ж. ‘тс, що ТРІСКАЛО’ 282;
БАБА[3] [ба́ба] -и, ж. *кулін.* ‘терта картопля, розлита на деко і спечена в печі’. Ба́бу тра́ба спечи́. 496 [Arkushyn].

We can show BABA[1] as a tree in accordance with the structure of vocabulary entry for a noun (with a headword as a root) and will see how many structural elements do not have their formal representations, either due to absence in general or through the “data compaction” of the articles in a paper dictionary:

БАБА[1]
(Phonetic transcription) [ба́ба]
(Gen. sing. form) -и
(Marker of gender for a noun) ж.
(Definition 1) жінка, яка має внуків [a woman, who has grandchildren]
(Stylistic marker 1) **N**/**A**
(Example 1) **N**/**A**
(Localization 1) (всі н. пп.) [all of the populated localities]
(Definition 2) будь-яка стара жінка [any old woman]
(Stylistic marker 2) **N**/**A**
(Example 2) **N**/**A**
(Localization 2) (всі н. пп.) [all of the populated localities]
(Definition 3) дружина [a wife]
(Stylistic marker 3) **N**/**A**
(Example 3) А твоя́ ба́ба вдо́ма? Вона́ ни бойи́ць’:а тако́го чолоўи́ка, йак ти? [Is your wife at home? Isn’t she afraid of such a husband as you?]
(Localization 3) 392, 11, 90, 342
(Definition 4) дівчина [girl]
(Stylistic marker 4) знев. [Disparagingly]
(Example 4) К’іровн’ічка хо́че нас в тре́т’ому посади́ти з баба́ми [Teacher wants us to sit with girls in the third form.]
(Localization 4) 392

It is evident from the dictionary entry that the example of the use of dialecticism appears only in entries 3 and 4 of BABA[1]. We do not have it for BABA[2]. A stylistic marker is present only in one of four values of BABA[1] and in BABA[3].

A tree structure can be much more difficult to produce in cases of several variants of the phonetic transcription:

ОБЛУД [Аркушин]
(Phonetic transcription 1) [облу́д]
(Gen. sing. form) -у
(Marker of gender for a noun) ч.
(Definition 1) тс, що БЛУД [the same as straying/wandering]
(Stylistic marker 1) **N**/**A**

(Example 1) В л'і́си облу́д бере́ [Oblud/Straying takes somebody in the forest. ∼ It is easy to lose one's way in the forest.]
(Localization 1) 129, 111
(Phonetic transcription 2) [о́блуд]
(Gen. sing. form) **N/A**
(Marker of gender for a noun) **N/A**
(Definition 1) **N/A**
(Stylistic marker 1) **N/A**
(Example 1) Тут о́блуд во́з'ме л'уди́ну [Here oblud/wandering will take a man.]
(Localization 1) 475, 432

The structure of the article varies according to the part of speech, the number of values, and the manner of pronunciation of the title lexeme, and accordingly has changing formal features. Formal features, in turn, can mutate both in accordance with structural changes and as a result of a change in the formal characteristics of other structural elements. In particular, the number of headword constituents influences the formal signs of such structural elements as phonetic transcription and Gen. sing. form (Gen. pl. – for Pluralia tantum). For example, the headword BAGNOVY LYS [bog fox] consists of two words, consequently, it has two phonetically transcribed words [багнови́й лис] and two parts of words for Gen. sing. form marking, which begin with a hyphen and end with one comma – -о́го -а (БАГНОВИЙ ЛИС [багнови́й лис] -о́го -а, ч. 'лис, який живе на болоті (в багнах); у нього шерсть темно-коричневого кольору' 122 [a fox that lives in a bog; its wool is dark-brown]) [Arkushin]. From a linguistic point of view this phenomenon is simply a linguistic fact, but at the level of the algorithmization of the addition of language data, this and other features cause additional steps in the working process.

When the researcher aims at processing many dictionaries in one chart, it is worth checking in advance how significant the difference in the structure of these dictionaries is. Compare the vocabulary entries for BABA in four sources (here presented in a single font face and size):

1. **БАБА**[1] род. одн. ы, род. мн. у, ув, ий, баб. Мати батька або матері. *Мойе баба і його баба пуд йен:им сонц'ом плат'а сушили.* Мд. *Така у нас чесна баба, так нас л'убит.* Тн. — Лпч., Вн, Нвц, Дрв, Ввр, Блк, Злт, ВБч,Яс, Он, Врц, Чн. Пестл. бабуна, бабунка. Дмш. 2. Стара жінка. *Д'ідовы без бабы нима порады.* Прч. *Та колис' і 'баба бы'ла 'д'івков.* Мд. *Гун'і сут' лем хи'ба у ста'рых ба'бу. Коло баб сто'йали и молôд'і 'жони.* Кй. — Збр, Пр, Ск, Кй, Лкц, Вн, Км, Кмн, Днл, Тх, Длв, Дв, Зч, Мд, Трв. Згруб. **бабера**. 3. Жінка взагалі, дружина. *Йа из с'войив 'бабов у'чора ц'ілий 'вич'ір быв.* Трн. *Майе вун премного ба'бий.* Трв. *Мойа баба шче лиш с'ім'нац:іт год'іў майе.* — Тн, Лпч, Вн, Трс, ЛПл. 4. Зневажл. (у порвнянні). *Ч'кода, же та'ку 'файну зарп'лату 'майе, а та'ка 'ходит', йак 'баба.* 5. (Зневажливо про чоловіка). *То'то тв'і чолов'ік не чолов'ік, але 'баба, ко'ли то'б'і дозво'л'айе 'вулиц'ами с'а ва'л'ати.* — Вн, Трс, Лкц, Врч. 7. Повитуха. *Йу'ришн'ій 'Мар'і ли'ше с'іл'с'ка 'баба помог'ла уро'дити д'і'тину.* Ізк. *Наша 'баба а'лумска 'дуже з'найе 'коло жун*

хо′дити. Вн. *У наш′ім сил′і ′добра ′баба, бо за нив ′д′іти ни вми′равут′.* Кв. — Ізк, Нвц, Бдв, Клч, Чн, Лв, Кв, Дв, Дмш, Вн. **АЛАЛ′С′КА (ва′лал′с′ка) ′баба, а′лумс′ка ′баба, вер′мец′ка ′баба, с′іл′с′ка ′баба, ′баба-пупо′р′ізка**. *Йак ув буде ой ′вармец′ка ′баба пома′гати, та ′буде ув малин′койе д′іт′âтко ревати.* Крц. *Йак жо′на хо′т′іла зл′е′чи, закликали ва′л′ас′ку ′бабу.* Зб. — Нвц, Лв. 8. Етн. Уявний образ злої потворної чаклунки (у сполученні з яга, босоркана). *Кой д′іт′ом ка′зати ′казку, та все бойатс′а лиш баби босоркані.* Вн. *Пуд мостом йе ′баба-йа′га.* Гтн. — Двг. 9. етн. Страшна, потворна стара жінка в народному повір′ї. *Хто перший раз дес′ йде, де шче не був, та ′кажуч, шо ′майе на до′роз′і ц′ул′увати страш′ну, шмар′каву та ґ′имбату ′бабу а дів′ки а′бо ′жони — боро′датого ′д′іда.* Трн, Лпч. **МАРТО′ВА БАБА**, етн. Міфічна істота, пов′язана з приходом весни. *Коли заг′р′ійе марта та тепло, та ′кажуч, що мартова баба з козами йде.* **СНІГОВА БАБА**. Зліплена зі снігу подоба людини. *Та′ку ′д′іти у вош′кол′і ′бабу зл′ї′пили, шо йшлам та напудилас′а.* Лпч. — Лкц, Тр. **СЛІПА БАБА**. Дитяча гра, піжмурки. Мж. **БАБА-КОРОЛАБА**, жарт. Дуже стара, худа баба. *Іде до нас ′баба-коро′ба.* Грш. (There is the first of 10 dictionary entries for BABA in [Grytsak]).

2. **Ба́бка**, -и, -ою(-ой). 1. Ворожбитка, знахарка, шептуха, чарівниця. [та|кий стаў ха|рошиў / йак с′ім |бабок поше^{и}п|тало] *Прос.* Див. ще **ве́дьма**. 2. Ріжки, грибковий паразит у колосках. [|жито по|йіла |баб∧ка] *Біллуц, Підг.* Див. ще **чорну́ха**. [SUSG].
3. БАБА: |*баба ро|зодрана* (34) ‘нецнотлива молода’. БАБИ |*баби* (60), *ба|би* (26, 52, 60, 62, 70, 76, 94, 101) 1. ‘присутні на весіллі жінки’ (101); 2. ‘свашки молодого / молодої’ (70, 94); 3. ‘жінки, що виготовляють весільне печиво’ (62, 94); 4. ‘жінки, які виготовляють весільні страви’ (26, 76); 5. ‘жінки, що виконують весільні пісні’ (52, 60, 94); 6. ‘жінки від молодої, що “продають” її придане родичам молодого’ (52); 7. ‘жінки, що заплітають косу молодій напередодні весілля’ (60). *ў|дал і ба|би* (94) ‘жінки, які виготовляють весільне печиво’. [Magrytska].
4. БАБА, наз. мн. ∼би /М-ць Мат. III, 33/, ∼би́ /К-в, О-в/, род. мн. ∼бі́ў /Б-те, Б-ля, Ж-ня, І-ї, К-в, Л-на В., Л-ць, П-п, П, Ш/, ∼б /Ж-ня/ 1. ‘стара жінка’; 2. ‘мати батька або матері’; 3. перен. ‘стара корова’ /Пас. 129/ ; 4. див. баби́ць /О-в/, пор. пол. baba; 5. ‘складова частина кросен’ /Ж-н Коб. 41, Я-ця/. - Ба́ба под′і́бна на пра́ник і вона́ вхо́дит у д′і́да: йе таки́й ко̀ло́к, шчо запиха́йес′і зве́рха ў ба́бу, шчоби́ шпанува́ло пр′а́жу /Я-ця/; заг. /Ж-н/. — Сто̀йі́т ба́ба у кут′і́ у зеле́н′ім кабат′і́, хто йіїї по̀ру́шит, гул′а́ти з неў му́сит (мітла) [Onyshkevych].

Every source has its own structural elements, located according to their own sequence, designed with an individual set of fonts and their descriptions. Experience shows that it is problematic to process two dictionaries by different authors using the same linguistic algorithm, oriented towards extracting the maximum detail for a vocabulary entry. It is therefore necessary to develop a separate algorithm for the processing of dictionary

entries for every new dictionary, or to make a stage-by-stage restructuring of linguistic information.

Therefore, three versions of the structure of electronic cards have arisen dependent upon the degree of complexity: simple, standard and extended. A standard version involves a restructuring of the dictionary entry according to the following levels: 1. REGISTER UNIT (WORD); 2. Interpretation 1; 2.1. Forms of word (through comma); 2.2. Grammatical descriptions; 2.3. Stylistic descriptions; 2.4. Examples; 2.5. Phraseology; 2.6. Localization; 2.7. Other. Preparation of the extended version of a dictionary entry involves the detailed realization of the explanatory part of a dictionary entry, addition of new dictionary material, and the marking of semantic (thesaural) relations between units of the database. At this stage of the work the simple version of card was selected as being the closest to the concept of card index and the fastest to fulfil practically, since it consists of two elements, marked out identically by all authors: register unit and teh semantic part of the dictionary entry. The subsequent processing of the card index is envisaged in order to broaden its structure (forming the extended electronic card) and to improve the system of search and sorting of linguistic data.

The other issue is the software representation of information in the database and the design of its interface. We made the structure of ELCIUD consist of six connected tables and a software interface (figure 1). The window of the card index contains a search engine, alphabetical index, editorial menu, menu for import and print, marker for number of register units, alphabet and index sorting form, window for a dictionary entry and form of dictionary entries sorted by their authors.

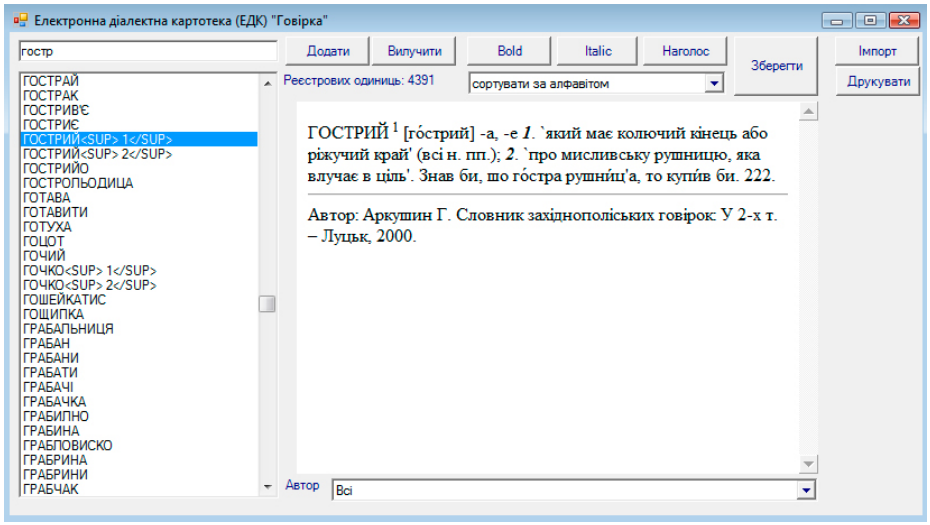


Fig. 1. The window of ELCIUD

ELCIUD is intended as an open system, so a system for addition of new information is necessary. Information addition can be achieved in two ways in ELCIUD:

1. addition of single linguistic units using the “Add” button of the menu
2. addition of data arrays by upload of two prepared files to the database through the menu “Import”.

These files must have a precise structure. The first file must contain details of the sources, written in accordance with the specially developed rules for preparation of the dictionary materials (complete and short names of a dictionary) and lists (localization of dialectal units after an author and after “Atlas of Ukrainian”, grammatical, stylistic markers, additional signs used for phonetic transcription etc). The second file consists of a dictionary, in which entries are given in a minimum (simple) version according to whether there is selected: 1. a register unit (word) or 2. the semantic part of a dictionary entry. The preparation of material in the format of standard and extended versions of vocabulary entry is also allowed. The difference between them is in the degree of detail of the dictionary material and these are not obligatory at this stage of the project development, unlike the first (simple) version.

The menu “Import” gives a possibility to add new dialectal units to the card index. It is recommended to upload preliminary prepared files to the database in the test mode at first (if there is an error in the structure of these files, the program will warn about it) and import of the tested file giving the numbers of added units. It is possible also to specify the necessary number of added headwords, giving the additional possibility of verification of data in the prepared file (figure 2).

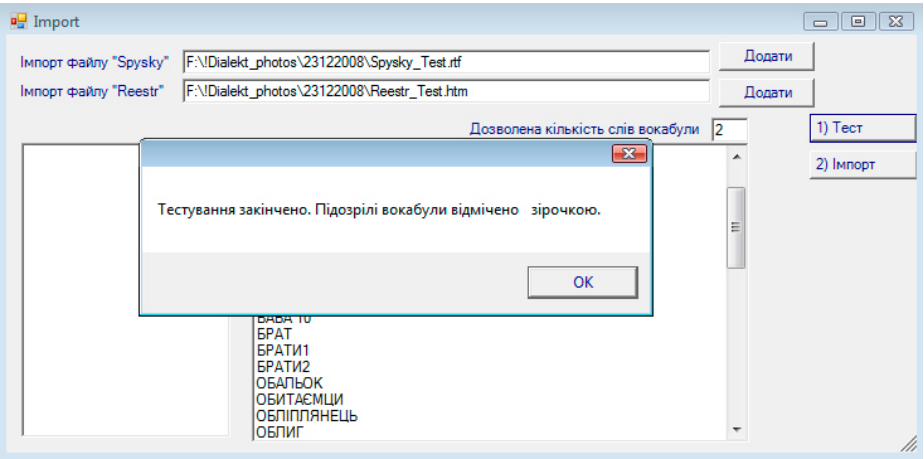


Fig. 2. The window for “Import” of the program ELCIUD after verification

However, in spite of the double system of verification for added files, there is still a need for the possibility of later correction of text and structural errors. An edit menu has been developed for this purpose, which enables working both at the level of register units and at the level of card texts. It is possible to correct text errors and edit font styles, and also unite dictionary entries which have become mistakenly disconnected.

The practical implementation of the system revealed problems with the upload of data. In particular, there were:

1. loss of accents and space characters
2. improper representation of the phonetic transcription signs
3. chopping of the dictionary entry

In order to avoid the first two problems it is necessary to save documents in .rtf format as web pages in Microsoft Word 2003. The third problem is solved by file testing described previously or by editing already added dictionary entries using the program menu (searching for bugs through index sorting and correction of contents using buttons "Додати"/"Видалити" ("Add"/"Delete").

The **problem of the complex localization mapping** of dialectal units also arose - the list of localizations and the card index exist separately in the database. Connection is complicated because one card unites several meanings and several briefly recorded localizations, respectively. A solution to this problem will be looked at in subsequent stages of work with the database, at the same time as the detailed realization of the dictionary entry.

The **problem of search and sorting** is related to the problem of localization mapping. Word search, author and word index sorting work in ELCIUD today. Enlargement of the system of search and sorting will be possible with the detailed realization of card structure.

In respect to the **fifth** problem, the prepared product (card index) is presented as a database in the Microsoft Access format and additional software (the program Electronic dialectal card index (ELCIUD), developed in Visual Studio in the C# programming language). Electronic cards are presented in electronic form only, but they can be copied, saved as a file and printed.

5 Prospects for the development of electronic card indexes

In the near future we plan to implement the following functions: possibility for printing several entries; author screening to prevent the addition of identical linguistic information; web-presentation of the card index, to produce the technical editing of all card indexes and to fill them with new register units.

The next and the most important task is the detailed realization of the card index for the general improvement of work with the material. This will involve extending the structure of the article, which will then consist of 21 elements. In accordance with such a structure, a new approach to the database and program is needed, allowing for transformation of the article in a new (extended) format. This will start a new stage in our work: a new project for the transformation of card indexes into the format of a complex electronic lexicographic system. This will relate the dialectal vocabulary array to the literary language array of the Modern Ukrainian Language Dictionary (an invariant field in the database) and will in future enable the structuring of linguistic data according to a concept criterion (fields of synonyms, antonyms, hyperonyms, and hyponyms).

Data structuring is also a method of perfecting the system for search and sorting, and for other valuable work with ELCIUD. It is the basis for the creation of database queries using one or several criteria, which is of great value for complex linguistic research. At this stage it is envisaged that there will be editing of the whole of the card index array, not only at the level of the content of entries, but also at the level of the connection and disconnection between register units.

6 Applications of ELCIUD

The project is intended for specialists and all those who are interested in the dialectal layer of the Ukrainian vocabulary. An electronic card index can be used as a basis for dialectological investigation, as an information system or for the purposes of education. In combination with other linguistic software systems, ELCIUD can serve as a basis for the analysis of the author's style and for complex research into Ukrainian at the different stages of its development.

7 Results

Within the framework of the project, the following tasks were performed:

1. theoretical principles of data preparation and the design of ELCIUD were drawn up; the key concepts are the system, structure and formalization at all stages of the work;
2. a database ELCIUD of six connected tables was produced in the Microsoft Access format;
3. an interface and software was developed in Visual Studio with the use of the C# programming language;
4. a menu for reading and editing the database was constructed;
5. the minimum sorting of dictionary information is implemented (alphabetical, index of headword and author sorting system);
6. we developed the possibility for enlargement of the database ELCIUD in two ways: addition of register units one by one through the menu and addition of files which contain the preliminary structured vocabulary entries;
7. a protocol was developed concerning ELCIUD data preparation by "virtual employees";
8. the test version of the web interface for ELCIUD was developed. An online version will be placed on the server of the Ukrainian Academy of Sciences or on the linguistic portal MOVA.info;
9. future transformation of ELCIUD to into a structured electronic lexicographic product is planned.

8 Sources

Arkushyn – Аркушин, Г. Л. *Словник західнополіських говірок.* У 2-х т. Луцьк 2000.

Grytsak – Грицак, М. А. *Словник українських закарпатських говірок.* – Машинопис. Зберігається в Інституті української мови НАН України.

Magrytska – Магрицька, І. *Словник весільної лексики українських східнослобожанських говірок (Луганська область).* Луганськ 2003.

Onyshkevych – Онишкевич, М. Й. *Словник бойківських говірок.* У 2-х т. Київ 1984.

SUSG – Глуховцева, К., Лєскова, В., Ніколаєнко, І., Тернавська, Т., Ужченко, В. *Словник українських східнослобожанських говірок.* Луганськ 2002.

References

[1] Гриценко, П. *Українська діалектна лексика: реальність і опис (деякі аспекти),* [in:] *Діалектна лексика: лексикологічний, лексикографічний та лінгвогеографічний аспекти,* Глухів 2005, с. 5–23.

[2] Жиго, П. *Опыт использования ЭВМ в обработке материалов ОЛА,* [in:] *Общеславянский лингвистический атлас. Материалы и исследования. 1991-1993. Сборник научных трудов,* Москва 1996, Вып. 21, с. 57–63.

[3] Крючкова, О. Ю. *Электронный корпус русской диалектной речи и принципы его разметки,* [in:] *Известия Саратовского университета. Серия "Филология. Журналистика",* т. 7, вып. 1, 2007, с. 30–34.

[4] Крючкова, О. Ю.; Гольдин, В. Е. *Текстовый диалектологический корпус как модель традиционной сельской коммуникации,* [in:] *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции "Диалог",* вып. 7(14), Москва 2008, с. 268–273.

[5] *Лексический атлас русских народных говоров. Материалы и исследования.* Санкт-Петербург 2005.

[6] Лесников, С. В. *Опыт информатизации диалектологических исследований.* Москва 1998.

[7] Тер-Аванесова, А. В.; Крылов, С. А. *Лексико-грамматические базы данных как инструмент диалектологического описания,* [in:] *Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог 2006",* Москва 2006, с. 493–498.

[8] Falińska, B. *Leksyka dotycząca hodowli na mapach Ogólnosłowiańskiego językowego atlasu.* Cz. 1-2. Białystok 2001.

[9] *Gwary polskie. Przewodnik multimedialny, pod red. H. Karaś,* [in:] `http://www.gwarypolskie.uw.edu.pl`

[10] Kąś, J. *Metody statystyczne w badaniach dialektologicznych,* [in:] Acta Universitatis Lodziensis. Folia Linguistica, vol. 12, 1994, pp. 117–123.

[11] Kretzschmar, W. *Art and Science in Computational Dialectology,* [in:] *Literary and Linguistic Computing,* vol. 21, n. 4, 2006, pp. 399–410.

[12] Kretzschmar, W. *Quantitative Areal Analysis of Dialect Features,* [in:] *Language Variation and Change,* vol. 8, n. 1, 1996, pp. 13–39.

[13] Nerbonne, J.; Kretzschmar, W. *Progress in Dialectometry: Toward Explanation,* [in:] *Literary and Linguistic Computing,* vol. 21, n. 4, 2006, pp. 387–397.

[14] Siruk, O. *Non-Technical Computer Thesaurus versus Specialized Computer Thesaurus,* [in:] *Metalanguage and Encoding Scheme Design for Digital Lexicography. Innovative Solutions for Lexical Entry Design in Slavic Lexicography. MONDILEX Third Open Workshop,* Bratislava, Ľ. Štúr Institute of Linguistics, 2009, pp. 177–182.

[15] Zarębina, M. *Słownictwo mieszkańców wsi Polski południowo-wschodniej (analiza statystyczna),* [in:] *Z polszczyzny historycznej i współczesnej,* Rzeszów 1997, pp. 35–50.

[16] `http://www.upei.ca/methodsxii/html/e_speakers_papers.html` (link created on February 18, 2009 to the materials of the seminars conducted during the work of the XII International conference Methods XII, held in August 2005 at the University of Moncton (Université de Moncton)).

Inflectional Entropy in Slovak

Adriana Hanulíková[1] and Doug. J. Davidson[2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
[2] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Abstract. Statistical measures of word frequency are used in psycholinguistic research to characterize the psychological organization of the mental lexicon, and the processes of retrieving, understanding, and learning words. More recently, researchers have calculated statistics from corpora to gain insights into processing of morphology, based on previous work on Serbian by A. Kostić and colleagues. One such statistical measure - the inflectional entropy - has been shown to explain processing costs in word recognition experiments. The inflectional entropy of a word form is the amount of information carried by that inflected form, relative to the statistical distribution of its inflectional paradigm. In this work, we investigate whether it is possible to calculate measures like inflectional entropy for Slovak using the Slovak National Corpus (SNK). This would allow us to compare Slovak with other Slavic languages such as Serbian. The results will be useful for a wide variety of psycholinguistic investigations of comprehension or production of Slovak.

1 Introduction

Many psycholinguistic investigations have shown that the probability of a word has a strong influence on measures of performance (for a recent review see Balota, Yap, & Cortese, 2006). This is true for a wide variety of tasks, such as word recognition, judgement tasks, or picture and word naming. For example, one of the most commonly-used tasks is the lexical decision task. In this task, the time it takes to judge whether a singly-presented word occurs in a language is measured. Response times in this task are faster for more common words relative to less common words (Whaley, 1978). Since a Slovak word like 'škola' (book) is used more often than a word like 'pštros' (ostrich), lexical decision times should be shorter for 'škola'.

For the purposes of psycholinguistic studies, the probability (Pr) of a word (w) is often approximated, as in Equation 1, by estimating its unigram frequency count $F(w)$ in a sample of text or speech of size N (Baayen, 2001). These counts are typically derived from non-annotated corpora, which do not provide information about grammatical classes or functions of the individual words.

$$Pr_w = F(w)/N \tag{1}$$

However, more recently researchers have incorporated variables related to morphosyntactic variation in the frequency estimates of words, based on annotated corpora (for review see Milin, Kuperman, Kostić, & Baayen, in press). This is especially important for Slavic languages, which have richer inflectional morphology than the

more-commonly studied West Germanic languages, and thus require more complex probability models. In particular, work on Serbian by A. Kostić and colleagues has been instrumental in demonstrating the influence of the inflectional form of a word on lexical decision performance. Since this framework is the point of departure for the present paper on Slovak, we will review some of their findings and conceptual distinctions here.

Kostić (1991, 1995) found that the relative frequency of an inflected form within a paradigm, as well as the number of grammatical functions or meanings of a word, was correlated positively with lexical decision times for Serbian nouns. Their measures were based on information theory, quantifying the amount of information that an inflectional suffix provides, relative to its paradigm. More recently, Moscoso del Prado Martín, Kostić, and Baayen (2004) found that lexical decision times for Dutch nouns were positively correlated with inflectional entropy. Inflectional entropy increases in a paradigm when there are more inflectional variants possible, and/or when the variants have similar probabilities. The key observation of this previous work is that the statistical distribution of word forms within an inflectional paradigm can be factored into two parts: The contribution provided by the stem, and the contribution conveyed by the exponent (i.e., suffix). This is illustrated below in Table 1, which shows a probability model for the Slovak feminine noun 'škola' (school), constructed in a similar way to Milin *et al.* (2009, in press). The columns provide information on the surface frequencies $F(w_e)$ (per million) and surface relative proportions $Pr_\pi(w_e) = F(w_e)/F(w)$, where $F(w)$ is the sum of all $F(w_e)$.

w_e	$F(w_e)$	$Pr_\pi(w_e)$	I_{w_e}	$F(e)$	$Pr_\pi(e)$	I_e
škol-∅	211	0.09	3.55	99396	0.11	3.25
škol-*a*	197	0.08	3.65	139469	0.15	2.76
škol-*u*	248	0.10	3.32	135748	0.14	2.80
škol-*i,y*	976	0.39	1.34	312564	0.33	1.59
škol-*e*	598	0.24	2.05	146867	0.16	2.68
škol-*ou*	66	0.03	5.23	68712	0.07	3.78
škol-*ám*	15	0.01	7.36	4890	0.01	7.59
škol-*ách*	146	0.06	4.09	17630	0.02	5.74
škol-*ami*	22	0.01	6.81	17576	0.02	5.75

Table 1. Probability distribution for the inflected noun *škola*.

The amount of information conveyed by the inflected words (w_e) and exponents (e) are calculated by applying the base -log2 transformation on the respective relative frequencies of the different exponents, and the relative frequencies of the inflected forms.

For example, the amount of information conveyed by the exponent 'u' (2.80) is calculated from the probability of the exponent $Pr_\pi(e)$

$$I_e = -\log_2 Pr_\pi(e) \tag{2}$$

where $e = u$ (0.1439), estimated from the frequency of the exponent $F(e)$ (135748) relative to the sum of the frequencies of the exponents in the paradigm (942852)

$$Pr_{\pi}(e) = \frac{F(e)}{\Sigma_e F(e)} \tag{3}$$

There are also other statistical measures which represent properties of the entire paradigm. The *entropy* of an inflectional paradigm, H, is calculated as

$$H = -\Sigma_e Pr_{\pi}(w_e) \log_2 Pr_{\pi}(w_e) \tag{4}$$

For the values shown in Table 1 for 'škola', this is calculated as: H('škola')$= -[0.0851 \times log_2 0.0851 \ldots 0.0089 \times log_2 0.0089]$, which amounts to 2.46. Informally, this index captures the degree to which the paradigm is unevenly distributed over the different forms.

In sum, these metrics characterize the contribution of stems and exponents to the probability that a word form will occur. These measures are made practically possible with the availability of relatively large morphosyntactically-annotated corpora such as the Slovak National Corpus (SNK).

Here we want to investigate whether it is possible to calculate inflectional entropy using the SNK, and if so, characterize how the results differ from previously reported results from Serbian. These comparisons would support future empirical research on word processing in Slovak, and help characterize differences between these two closely related languages.

Number	Case	Serbian	Slovak
Singular	Nominative	planin-*a*	planin-*a*
	Genitive	planin-*e*	planin-*y*
	Dative	planin-*i*	planin-*e*
	Accusative	planin-*u*	planin-*u*
	Instrumental	planin-*om*	planin-*ou*
	Locative	planin-*i*	planin-*e*
Plural	Nominative	planin-*e*	planin-*y*
	Genitive	planin-*a*	planín-∅
	Dative	planin-*ama*	planin-*ám*
	Accusative	planin-*e*	planin-*y*
	Instrumental	planin-*ama*	planin-*ami*
	Locative	planin-*ama*	planin-*ách*

Table 2. Slovak and Serbian regular feminine inflectional exponents, illustrated with the noun 'planina' (meaning mountain in Serbian and plain in Slovak).

Despite the differences between surface exponents used in Serbian and Slovak (see Table 2 above for an example), there are many similarities between the morphosyntactic systems of Slovak and Serbian. Both languages have relatively complex inflectional systems, in which nouns are marked for number (singular and plural) and grammatical case

(nominative, genitive, accusative, dative, instrumental, locative; the vocative is archaic in Slovak and its status is disputed in Serb). In addition, the inflectional endings depend on the gender of the noun (feminine, masculine, neuter) and the inflectional class.

Given such similarities, we would expect that statistical distribution of the Serbian and Slovak terms would be similar. If we take the example of a base-level term used in Milin *et al.*, such as 'žena' (woman), we should observe a similar statistical distribution as their Slovak counterpart 'žena', because they would be expected to have a similar distribution of grammatical functions and meanings. If this is the case for most of the terms in Slovak, then many of the psycholinguistic results obtained from the study of Serbian should also generalize to Slovak.

On the other hand, there might be some reasons to expect differences between these (and also other Slavic) languages. First, some of the basic-level terms in the two languages have different meanings, gender or inflectional class. For example, the primary meaning of 'planina' (mountain in Serbian) does not correspond to the same meaning as its Slovak counterpart 'planina' (plain in Slovak). Second, the statistical estimates for Serbian are based on a *sample* of text, as is the case with all statistical parameter estimates. It may be the case that the parameter estimates for a given measure like inflectional entropy will be conditioned on the data source. This would suggest that the Slovak and Serbian parameter estimates could be different, either due to real differences in the usage of the two languages, or to differences in the samples used to estimate the parameters.

We hypothesized that the factors governing the paradigm distribution of nouns in Slovak and Serbian would be similar. We predicted that the measures of inflectional entropy and paradigm entropy of Slovak and Serb would therefore also be similar.

2 Method

For a global comparison with Serbian results, we created two figures as in Milin *et al.* (2009:55). We made a query from the SNK for all feminine and masculine nouns in all respective cases and numbers. We then extracted statistical information for all feminine exponents. Masculine nouns were not further analyzed. Milin and colleagues focused on dominant regular inflectional subclasses in their paper; we consider all feminine exponents. Note that ⟨y, i⟩ exponents were not computed separately, since in modern Slovak they both express the same phoneme /i/. The function of ⟨y⟩ is to indicate that the preceding sound is not palatalized.

For the comparison of inflectional entropy between the two languages, we selected words from the word list provided in Milin *et al.* (2009) for which there was (almost) complete form overlap with their Slovak counterparts, and used these for the query in the manually morphologically annotated subcorpus *r-mak-3.0* from the SNK. For the analysis, we used only those words that were present two or more times in the SNK sample, and we did not include diminutives. The frequencies and relative frequencies of inflected variants and inflectional exponents were computed in the same way as in Milin *et al.* (2009) and as described earlier in the Introduction.

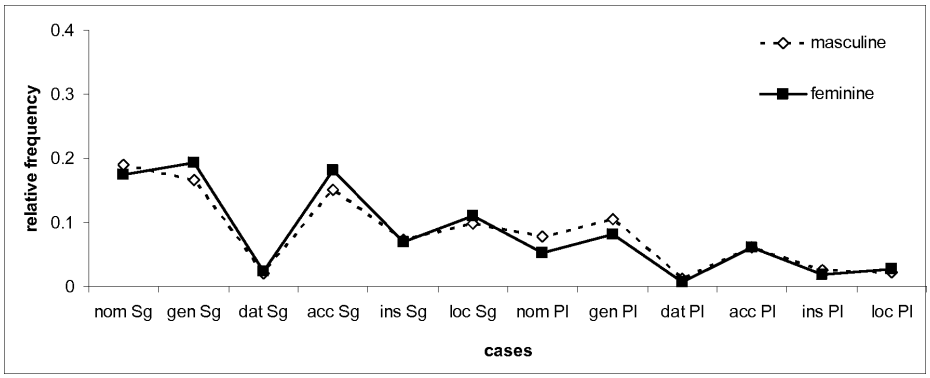


Fig. 1. The relative frequencies of feminine and masculine nouns for Slovak according to case and number.

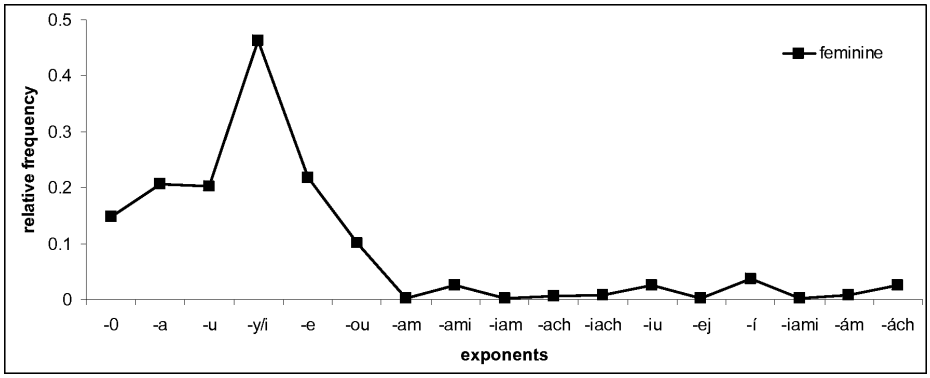


Fig. 2. Relative frequency of feminine nouns in Slovak according to inflectional suffix.

3 Results and discussion

Figure 1 shows for each case-number combination the distribution of relative frequencies within each inflectional class (here, the masculine and feminine nouns). Except for the values of relative frequencies, the picture is almost identical to the Serbian results. This is a good example of how different corpora can still be representative with respect to morphological aspects of language use, irrespective of whether it is of a smaller or larger size. Figure 2 plots the relative frequency of individual exponents within the feminine inflectional classes. These are also considerably similar to Serbian.

Now we turn to the question, whether the inflectional entropy of individual cases is comparable as well. Table 3 shows the inflectional entropy, H, calculated for the words we selected from the Serbian lists. The average entropy for Slovak ($\mu = 1.70$), in this sample, was less than Serbian ($\mu = 2.11$), $t(18) = 2.011$, $p = 0.059$. The correlation between the two samples was relatively low, $r = 0.2$. This result would suggest that the deviation from the paradigm pattern is, on average, greater for Serbian than for Slovak.

Slovak	H	Serbian	H
kniha	2.63	knjiga	2.17
rieka	2.28	reka	2.22
búrka	1.30	bura	2.23
tráva	1.52	trava	2.23
brigáda	0.65	brigada	1.89
fabrika	0.86	fabrika	2.12
škola	2.46	škola	2.20
náuka	0.88	nauka	1.98
ruža	1.24	ruža	1.90
stanica	1.72	stanica	2.05
ulica	3.04	ulica	2.39
dolina	0.59	dolina	2.43
duša	2.36	duša	2.28
ryba	1.71	riba	1.79
sila	3.27	sila	2.03
potreba	2.74	potreba	2.13
vŕba	0.24	vrba	1.86
hlava	0.80	glava	2.34
hviezda	2.01	zvezda	1.83

Table 3. Comparison of Slovak and Serbian word pairs.

This result suggests that despite the similarities between Serbian and Slovak, their inflectional entropy differs. However, several caveats should be kept in mind. This comparison was based on a relatively limited number of words, and in order to maintain strict comparability, we only examined words with overlapping surface forms. Despite this overlap, preferences for certain terms, or differences in meaning in the respective languages, could lead to differences in the frequencies of some terms. Future work could examine larger samples, and other inflectional classes.

Despite the small sample, the results offer some suggestion that individual measures of entropy are needed for each language, even for languages as typologically similar as Serbian and Slovak. In practical terms, it appears that the use of morphologically-annotated corpora are very helpful for calculating these measures for each language. A useful framework for future comparisons of Slavic languages (or other languages that have similar inflectional classes) might include measures like inflectional entropy in order to guage the similarties and differences between languages.

4 Summary

In this paper we have described how inflectional entropy can be estimated from the Slovak National Corpus. The obtained estimates were compared to results reported previously for Serbian. The results showed that overall, the distribution of feminine and masculine inflected nouns (grouped according to case and number) is almost identical for both languages. The comparison of relative frequencies for feminine nouns, grouped

by inflectional suffixes, showed a considerable amount of similarity with Serbian, despite the differences in suffix forms. Given this outcome, we expected inflectional entropy measures for a selected number of Slovak and Serbian (high frequency) nouns to be comparable. However, the results showed that the estimates differ. This implies that morphologically-annotated corpora could be very useful for cross-linguistic comparisons.

5 Acknowledgements

We would like to thank the producers of the Slovak National Corpus for making their work widely available. This work was made possible by the support of the Max Planck Gesellschaft. Both authors contributed equally.

References

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (A travel update). In M. Traxler and M. Gernsbacher (Eds.) *Handbook of Psycholinguistics, 2nd Edition*. Pp. 285–375. Amsterdam: Elsevier.

Kostić, A. (1991). Informational approach to the processing of inflected morphology: Standard data reconsidered. *Psychological Resarch, 53*, 62–70.

Kostić, A. (1995). Informational load constraints on processing inflected morphology. In L. B. Feldman (Ed.) *Morphological Aspects of Language Processing*. Pp. 317–344. New Jersey: Lawrence Erlbaum Inc. Publishers.

Milin, P., Kuperman, V., Kostić, A., & Baayen, R. H. (in press). Words and paradigms bit by bit: An information-theoretic approach to the processing of inflection and derivation. In J.P. Blevins & J. Blevins (Eds.), *Analogy in grammar: form and acquisition*. Oxford University Press: Oxford.

Milin, P., Durdevic, D. F., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on recognition: Evidence from Serbian. *Journal of Memory and Language, 60*, 50–64.

Moscosos del Prado Martín, F., Kostić, A., Baayen, R. H. (2004). Putting the bits together: An informational theoretical perspective on morphological processing. *Cognition, 94*, 1–18.

Slovenský národný korpus, – r-mak-3.0. Bratislava: Jazykovedný ústav. Ľ. Štúra SAV 2008. `http://korpus.juls.savba.sk`.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*, 143–154.

Exploring Derivational Relations in Czech with the Deriv Tool*

Dana Hlaváčková[1], Klára Osolsobě[2], Karel Pala[1], and Pavel Šmerk[1]

[1] Faculty of Informatics, Masaryk University
Botanická 68a, 60200 Brno, Czech Republic
{hlavackova,pala,smerk}@fi.muni.cz
[2] Faculty of Arts, Masaryk University
Arna Nováka 1, 602 00 Brno, Czech Republic
osolsobe@phil.muni.cz

Abstract. The aim of this paper is to present a tool for testing automatic word derivation for Czech. The derivation of some word formation types in Czech is in high degree regular. It can be described by formal rules. A new version of the web interface Deriv working with the morphological analyzer ajka enables us to formulate more complex word formation rules and to test more complicated cases of derivational relations. The second issue touched extensively in the paper are the types of derivational relations and their semantic classification. We have proposed 14 semantic classes for suffixes and 11 for prefixes. The tool Deriv helps considerably in establishing semantics of the derivational relations.

1 Introduction

In the paper we present the recent results obtained in the area of automatic processing of Czech morphology (inflectional and derivational) in cooperation between the Department of Czech Language, Faculty of Arts, Masaryk University (K. Osolsobě) and the Centre for Natural Language Processing, Faculty of Informatics of the same university (D. Hlaváčková, K. Pala, P. Šmerk).

In section 2 we briefly characterize existing tools for automatic processing of the formal morphology of Czech (inflection) as well as the derivational morphology (word formation). In the following section we explain how changes on the formal level correspond to semantic relations between related words (motivation relations). In section 4 we show how the relations between word base and derived word can be described formally. We also offer some examples showing how the particular rules can be formulated and used in the new version of the software tool Deriv which allows us to test these derivational rules on a large machine dictionary of Czech word forms. In the next section we present some obtained results, especially how the new features in the tool Deriv allows to formulate more general rules without a significant loss of precision. Finally, in section 6 we summarize the results obtained so far in the course of testing the software tool Deriv and outline the further possibilities of its use.

* This work has been partly supported by the Ministry of Education, Youth and Sports within the National Research Programme II project 2C06009 and project LC536.

2 Tools for (Czech) morphology

We work with two software tools: the first one is earlier developed morphological analyzer `ajka` [13] and the second is newer web derivational interface Deriv [14] which uses `ajka` and its data (however the tool itself is language independent).

2.1 Morphological analyzer `ajka`

The first of our tools that is able to process Czech word forms (both recognition and generating) as well as some regular derivations is the morphological analyzer `ajka`[3]. It was developed in the Centre for Natural Language Processing at the Faculty of Informatics, Masaryk University. The tool works with the formal description of the Czech inflectional paradigms that was developed by Osolsobě [8] and with the list of Czech stems (approx. 400 000 items).

2.2 Interface Deriv

The second tool, which we are going to present in detail, is a web interface Deriv[4]. It is an interactive tool able to process derivational relations in Czech. According to the defined rules it can generate lists of words from the Czech word-form list (more than 60 000 000 items). It is based on the assumption that it is possible to extract n-tuples of the items from the wordlist of `ajka` (or from corpus) that meet two following requirements:

- all members can be found in the wordlist;
- each member of the n-tuple meets the underlying hypothesis.

The list of n-tuples generated automatically in this way is subsequently manually checked and the list of correctly derived n-tuples and the list of the exceptions (over-generated cases) is obtained. The Deriv tool works in three steps:

- by using a grammatical tag and regular expression both appropriate strings and substitutive rule are defined;
- the respective pairs base form/derived form are automatically generated;
- the obtained list of the n-tuples is further processed manually.

Incorrectly derived word-form are also collected as negative examples for the next iteration of the more complex rules.

3 The types of the derivation relations and their semantics

The semantic relations (traditionally characterized as motivation) can be captured by formal rules based on regularities on the level of graphical realisation of phonemes, combining affixes and parts of speech. They exploit Czech traditional word formation grammar [3], and semantic classifications worked out for Czech WordNet and lexical database VerbaLex [5].

[3] More information at `http://nlp.fi.muni.cz/projekty/ajka`, the interactive version is accessible at `http://nlp.fi.muni.cz/projekty/wwwajka`.

[4] More information at `http://deb.fi.muni.cz/deriv`.

3.1 Semantics of the derivational relations

In our view, there is a need for a more detailed and better semantic specification of the derivational relations (D-relations) in comparison with the traditional labels that we consider rather broad and too general.

On the ground of the previous research [5] we offer a collection of the 15 derivational relations (see also [2] and [6]). We try to capture the main relations and leave aside either the marginal ones or the ones that are not using strictly suffixation and prefixation, i.e. compounds.

They can be tentatively grouped in the following way:

- role D-relations in which a relation between two POS expresses a semantic role, e.g. Agentive, Patient, Instrument, Location, etc.;
- gender, diminutive and augmentative D-relations;
- D-relations denoting various kinds of properties, some of them can be considered deverbative (first member of the pair is always a verb) and the remaining ones exist between nouns and adjectives (also possessives) and adjectives and adverbs;
- prefix D-relations consisting only of the pairs verb – prefixed verb, their meanings are partly related to the verb classes, e.g. verbs of motion, verbs of drinking and eating, or verbs of weather, etc. This area represents a challenge for further research.

From the first 14 D-relations using suffixation the basic and most productive ones have already been integrated into Czech morphological analyzer `ajka` together with their semantic labels. The 15th one is, in fact, a complex group of the 11 relations existing between verbs and using prefixation only. They are mentioned briefly below.

The indicated types of the D-relations can be further characterized in the following way:

- D-relations expressing semantic roles similar to deep cases, they exist between verb – noun pairs and use suffixation (numbers in the brackets denote frequency of the pair in the corpus SYN2000[5], which is part of Czech National Corpus [1]):
  - der-agentive: *učit – učitel* (8639 – 9924), *teach – teacher* (pair verb – noun);
  - der-patient: *trestat – trestanec* (1636 – 301), *punish – convict* (pair verb – noun);
  - der-instrument: *ukazovat – ukazovátko* (11 831 – 61), *point – pointer, fescue* (pair verb – noun);
  - der-location: *letět – letiště* (3213 – 6636), *fly – airport* (pair verb – noun);
- D-relations denoting gender derivating of feminine nouns from their masculine counterparts and and D-relations expressing diminutivity and augmentation, they exist between noun – noun pairs. In Czech deminutives occur also in triples (suffixation):
  - der-gender: *student – studentka* (15 608 – 1260), *student – she-student* (pair noun – noun);

[5] A detailed information on the structure of corpus SYN2000 can be found at `http://www.korpus.cz/english/struktura.php`

  - der-dimin: *dům – domek – domeček* (46 485 – 5118 – 606), *house – small house, very small house or house one likes* (triple noun – noun – noun);
  - der-augm: *dub – dubisko* (752 – 24), *oak tree – huge and strong oak tree* (pair noun – noun);
- D-relations denoting action and property (suffixation) – their typical feature is that both members of the relation refer to the same meaning and they differ only in the part of speech, for instance, in the pair *učit – učení (teach – teaching)* the action is denoted by a verb and deverbative noun, respectively. In the other pairs both members denote property and the difference consists again only in the part of speech:
  - deriv-action: *učit – učení* (8639 – 2877), *teach – teaching* (pair verb – deverbative noun);
  - deriv-property: *učit – učený* (8639 – 409) *teach – learned* (pair verb – deverbative adjective);
  - deriv-property: *učený – učenost* (409 – 63) *learned – learnedness* (pair adjective – noun);
  - deriv-property: *učený – učeně* (409 – 33) *learned – learnedly* (pair adjective – adverb);
  - deriv-property-possessive: *učitel – učitelův* (9924 – 56) *teacher – teacher's* (pair noun – adjective);
- D-relations exploiting prefixation represent a separate complex group in which we observe only the pair verb – verb. They express meanings depending heavily on the meanings of the verb stems they occur with. They denote a number of different semantic relations such as various sorts of motion, time and location (see below), intensity of action, inchoativity, iterativity, additivity, distributive action, obligation, result and possibly some others. Here we offer their preliminary classification which obviously requires a further investigation. It is based on the list of 14 Czech basic (primary) prefixes that can be usually found in Czech grammars [12]: *do- (to), na- (on, at), nad- (above, up), od- (from, away), pro- (for, because), při- (by, at), pře- (over), roz- (over), s-/se- (with, by), u- (at, near), v-/ve- (in, up), vy- (out, off), z-/ze- (of, off), za- (over, behind)*.

  Verbal D-relations using prefixes can be further characterized:

- motionI: deriv-mot-to: motion to the point or place, e. g. *jít – přijít, (go – come*), *letět – přiletět (fly – arrive by plane)*;
- motionII: deriv-mot-to-iter: iterative, repeating motion to a point or place, e. g. *přicházet – přicházívat (be coming – be coming repeatedly)*;
- motionI: deriv-mot-from: motion from a point or place, e. g. *jít – odejít (go/walk – leave by going/walking)*;
- motionI: deriv-mot-from-iter: iterative motion from a point or place, e. g. *odcházet – odcházívat (leave by walking – leaving repeatedly)*;
- motionI: deriv-mot-over: motion across a point or place, e. g. *brodit – přebrodit (wade – wade through)*;
- motionII: deriv-mot-under: motion under a point or place, e. g. *letět – podletět (fly – fly under)*;

- timeI: deriv-compl-act-: to complete an action (with regard to any verb), e.g. *letět – doletět (fly – finish flying)*;
- timeII: deriv-t-act-iter: to complete an action iteratively, e.g. *tancovat – dotancovat – dotancovávat (finish dancing – finish dancing repeatedly)*;
- obligation: deriv-oblig: to perform an action as an obligation *pracovat – odpracovat (work – work off)*;
- additivity: deriv-addit: action expressing adding *koupit – přikoupit (buy – buy more)*;
- distributivity: deriv-distrib: to perform an action in a distributed way *ztratit – poztrácet (lose – lose little by little)*, i.e. lose successively particular objects one after another;
- result: deriv-result: to perform an action with its result *vařit – vyvařit (boil – boil away)*;
- high intensity: deriv-high-intens: to perform an action more intensively *vařit – navařit (cook – cook a lot of sth)*;
- low intensity: deriv-low-intens: to perform an action with low intensity *pracovat – popracovat (work – work a little, for a while)*.

The list is preliminary and in our view it includes just the main and most typical meanings expressed by the primary prefixes *do- (to), od- (from), na- (on, to), po- (after), pod- (under), pře- (over, across), při- (to), vy- (out, from)*. We give examples of all meanings mentioned above but not necessarily with all prefixes. It can be seen that meanings related to motion and time can be further subclassified but we certainly have not captured all that can be related to the verbs of motion. In this respect the more detailed investigation is necessary. We also consider only the verbs of motion with two arguments, i. e. verbs with an Agent causing motion to a Location. Verbs with an Agent, moved Object, e.g. *nést knihu domů – přinést knihu domů (carry the book home – fetch the book home)* and Location (home) still have to wait for the more detailed analysis. In the list above we have included iterativity relation as well because of its regularity in Czech though iteratives are not derived with prefixes but with alternations in the stems using infixes *-áva-, -íva-, -ova-*. On one hand, including the iterativity relation may seem to complicate the description but on the other, its regularity allows us to handle it almost automatically. It has to be remarked that the iterativity relation is semantically close to the aspect relation, that is why some authors speak about third aspect though iteratives are imperfective by definition.

We are well aware that the D-relations using prefixation can be organized differently as it is typical for any kind of semantic classification. We are attempting to find one of the possible solutions.

The need for describing semantics of the D-relations comes from the fact that D-relations are directly accessible for language users and make them able to understand meaning relations existing in the the text. Their formal description is a necessary condition for implementing more intelligent semantic search in the applications like Semantic Web.

4 Word derivation rules for Deriv

Word derivation, i.e. forming new words from the corresponding word bases (words), can be formally described as an operation over strings of the characters (lemma, word

form, morphological tag). The derivational relations are typically regular and can be described as formal rules for the Deriv interface. This allows us to reduce the machine wordlist as well as to minimize the list of roots.

The very first versions of the tool Deriv only allowed formulation of rules based on simple substitution of a final (or initial) substring of characters. Later the support for regular expressions was added, so that the rules could describe also some simple alternations in a more general manner.

4.1 New features of Deriv

To allow construction of even more general rules we have implemented the following two features:

- the rules now can describe relations not only between lemmata but also between arbitrary word forms (previous versions of Deriv were able to work with lemmata only);
- we have added three general rules to describe alternations which cannot be directly expressed by means of regular expression substitutions:
  - palatalization of the final consonant;
  - shortening of the last vocal;
  - shortening of the last but one vocal.

Moreover, to allow further simplification of the rules we have added two another features:

- utilisation of regular expressions[6],of the programming language Perl 5.10, especialy of the construct \K which separates what is to be replaced and what have to precede (in previous versions whole match was replaced);
- introduction of four shorthands $C, $V, $L, and $S for character classes “consonant”, “vowel”, “long vowel”, and “short vowel” respectively.

4.2 Structure of the rules for Deriv

The rules for Deriv have four parts:

- regular expression which describes the tag of the first word of the relation;
- regular expression which describes the tag of the second word of the relation;
- regular expression which describes which part of the first word is to be replaced;
- the replacing string.

The regular expressions use a syntax of Perl 5.10 programming language regular expressions. The tags uses an atribut-value notation in which e.g. k1 stands for a noun, k2 for an adjective, k5 for a verb, gM for a masculine animate, gI for a masculine inanimate, gF for a feminine, gN for a neutral, nS for a singular noun, c1 for a nominative, c2 for a genitive, and mF for an infinitive[7]. The application selects words with the tag

[6] For full documentation see `http://perldoc.perl.org/perlre.html`

[7] For full documentation see `http://nlp.fi.muni.cz/projekty/ajka/tags.pdf` (in Czech).

matching the respective regular expression and then among these words it searches for pairs in which the substitution of a string to be replaced in the first word of the pair for the replacing string produces the second word of the pair. Only the pairs which meet all conditions are returned to the user.

4.3 Some examples of the rules for Deriv

At first, an example of a simple rule – a particular class of adjectives can be derived from the animate nouns with the suffix *-í*. The rule for automatic processing of pairs like *pták/ptačí (bird/of a bird)* then is:

tag which selects the 1st word: k1gMnSc1	string to be replaced: ák$
tag which selects the 2nd word: k2.*gMnSc1d1	replacing string: ačí

The rules are usually writen on one line in the form

ák$/k1gMnSc1 > ačí/k2.*gMnSc1d1

Examples of more complex rules A rule for an automatic processing of pairs like *řídit/řidič (drive/driver)* can be formulated as follows (note the (.[aeěi]) and $1 constructs – whatever matching is between *í* and *t* in the first word, it is left in the second word):

í(.[aeěi])t$/k5.*mF > i$1č/k1gMnSc1

Another interesting example is a rule for an automatic processing of pairs like Altaj/altajský, where a change of capitalization of the first letter has to be described:

^([[:upper:]].*)/k1gInSc1 > \l$1ský/k2.*gMnSc1d1

Examples of rules exploiting new features of Deriv The rule

LC\K[ay]$/k1g[FM]nSc2 > í/k2.*gMnSc1d1 /Mk

is a broad generalisation of the rule for pairs like *pták/ptačí* described above. /Mk at the end of rule denotes an application of two of the general rules “palatalization of the final consonant” and “shortening of the last vocal”, so that not only *á* and *k* will be shortened and palatalized respectively, but also e.g. *ů* and *h* in *bůh/boží (god/of a god)* etc. The first regular expression says that the first word of a pair has to end with *a* or *y* which has to be preceded by a sequence of any long vowel and a consonant, but only the final *a* or *y* will be replaced (because of \K in between). Moreover, note that the rule derives from genitive forms (c2 in the first tag), not from nominative forms (lemmata).

Another example is a rule for an automatic processing of pairs like *uchvátit/uchvatitel (usurp/usurper)*:

LC+$Vt\K$/k5.*mF > el/k1gMnSc1 /K

where /K at the end of the rule denotes an application of the general rule “shortening of the last but one vocal”. Note the \K at the end of the first regular expression – it denotes that only the words which ends with a long vowel, one or more consonants, any vowel and *t* will be retrieved, but nothing will be replaced, so that only the addition of the string *el* will occur.

5 Results

In the Table 1 we show how the new features of Deriv make possible to formulate derivational rules more generally. The table compares sets of rules for four derivational types which we studied in our previous work [4] (where one can found all particular rules together with numbers of found pairs and over-generations for each rule and also examples of the most of over-generations). The new rule sets have been constructed in such a way that they find all correct pairs which are found by the original rule sets.

derivational type *example*	old version of Deriv			new version of Deriv		
	rules	pairs	over-generation	rules	pairs	over-generation
agentive (*-tel*) *užívat/uživatel (use/user)*	22	907	14 (1.5 %)	12	912	15 (1.6 %)
agentive (*-č*) *loupit/lupič (rob/robber)*	14	910	16 (1.8 %)	10	917	16 (1.7 %)
purpose adjective (*-í*) *krýt/krycí (cover [v/adj])*	5	1746	3 (0.2 %)	1	1746	1 (0.1 %)
generic possesives (*-í*) *husa/husí (goose [n/adj])*	23	600	165 (27.5 %)	6	481	18 (3.7 %)[8]

Table 1. Comparison of old and new version of Deriv

In all cases there is a significant decrease of number of rules without any substantial increase of over-generation. Also in all cases the new, more general rules found some pairs which had been omitted before (mostly because of uniqueness of its alternations).

The Table 2 shows results of automatic processing of formally defined pairs verb – agentive noun in Deriv and their manual classification[9]. The figures for the first two suffixes are the same as in the Table 1. But because the new features of Deriv are really recent, we did not manage to work over the rule sets for all derivational relations for the time being. Therefore to preserve mutual comparability of the figures in this table we have to use older figures in all cases.

These results document an interesting fact that if the agentive nouns are derived from a verbal stem (root + thematic infix, suffixes *-tel* and *-č*), the rate of over-generation is significantly lower in comparison to cases where the agentive nouns are derived directly from the root (suffixes *-ce*, *-čí* and *-ec*), even though these derivational types are described by larger sets of rules. In other words, the derivation with the former suffixes appears to be much more regular than the derivation with the latter ones.

[8] This outstanding decrease of a rate of over-generation is caused by splitting feminines according to animateness: it is very essential information in the context of these rules as the generic possesives can be derived only from the animate nouns.

[9] More information on the derivation with suffixes *-tel* and *-č* can be found in [4, 9, 10], on the suffixes *-ce* and *-ec* in [7] and on the suffix *-čí* in [11].

suffix	example	rules	pairs	over-generation
-tel	*užívat/uživatel (use/user)*	22	907	14 (1.6 %)
-č	*loupit/lupič (rob/robber)*	14	910	16 (1.8 %)
-ce	*soudit/soudce (judge/judge)*	74	326	51 (15.6 %)
-čí	*mluvit/mluvčí (speak/speaker)*	18	42	11 (26.2 %)
-ec	*jezdit/jezdec (ride/rider)*	28	121	45 (37.2 %)

Table 2. Results of processing of verb – agentive noun pairs

6 Conclusions

We have studied regularities of the Czech word formation and using the improved version of the tool Deriv we have formulated and evaluated formal derivation rules for different derivational relations. We have also paid attention to the semantics of the D-relations – the final task is to label the derivation rules also semantically and in this way to make them usable in various NLP applications. Though a number of the individual types have not been processed automatically yet, the presented results convincingly show that Deriv allows to explore potential and limits of the formal description of the derivational processes. The results can be used in NLP applications where semantic search becomes the aim, but also e.g. for optimisation of existing software tools, especially the morphological analyser `ajka` and its stem list.

References

[1] (2000). Czech National Corpus – SYN2000. Institute of the Czech National Corpus, Praha. Accessible at: `http://www.korpus.cz`.

[2] Azarova, I. V. (2008). Derivational semantic relations in RussNet. *An oral presentation at the 4th Global Wordnet Conference*. Szegéd, Hungary.

[3] Dokulil, M. (1962). *Tvoření slov v češtině I (Word Derivation in Czech I)*. Nakladatelství ČSAV, Praha.

[4] Hlaváčková, D., Osolsobě, K., Pala, K., and Šmerk, P. (2009). Relations between formal and derivational morphology in Czech. In *Proceedings of Czech in Formal Grammar 2009 Conference*. Peter Lang Publishing Group. In print.

[5] Hlaváčková, D. and Pala, K. (2007). Derivational Relations in Czech WordNet. In *Proceedings of the Workshop BSNLP*, pages 75–81, Prague. ACL.

[6] Koeva, S., Krsteva, C., and Vitas, D. (2008). Morpho-semantic Relations in WordNet – a Case Study for Two Slavic Languages. In *Proceedings of 4th Global WordNet Conference*, pages 239–253, Szegéd, Hungary.

[7] Kolářová, Z. (2009). Možnosti a meze automatické derivace – počítačové zpracování deverbativ na -ce a -ec *(Potential and Limits of an Automatic Derivation – Computer Processed Deverbatives with -ce and -ec Endings)*. Master's thesis, Faculty of Arts, Masaryk University, Brno.

[8] Osolsobě, K. (1996). *Algoritmický popis české formální morfologie a strojový slovník češtiny (Algorithmic Description of Czech Inflectional Morphology and Czech Machine Stem List)*. PhD thesis, Faculty of Arts, Masaryk University, Brno.

[9] Osolsobě, K. (2008a). Formální pravidla derivace deverbativ na -č *(Formal Rules for Derivation of Deverbatives ending with -č)*. *SPFFMU A*, 56:121–135. Faculty of Arts, Masaryk University, Brno.

[10] Osolsobě, K. (2008b). Propria (příjmení na -č) – problém automatické morfologické analýzy *(Family Names with -č ending – a problem of automatical morphological analysis)*. In Čornejová, M. and Kosek, P., editors, *Jazyk a jeho proměny (Language an it's Metamorphosis). Prof. Janě Pleskalové k životnímu jubileu*, pages 205–216, Brno. Host.

[11] Osolsobě, K. (2009). Deriváty na -čí: gramatika, slovník a korpus *(Derivates with -čí: grammar, dictionary and corpus)*. *SPFFMU A*, 57. Faculty of Arts, Masaryk University, Brno. In print.

[12] Petr, J. (1986). *Mluvnice češtiny I (Grammar of Czech I)*. Academia, Praha.

[13] Sedláček, R. (2004). *Morphematic analyser for Czech*. PhD thesis, Faculty of Informatics, Masaryk University, Brno.

[14] Šmerk, P. (2009). *Deriv*. Web application interface (in Czech), accessible at: `http://deb.fi.muni.cz/deriv`.

On Epistemicity, Grammatical Person and Speaker Deixis in Polish (Based on the Polish National Corpus)

Łukasz Jędrzejowski

Interdisciplinary Center "European Languages: Structures – Development - Comparison"
Free University of Berlin, Germany

Abstract. Examining data from the Polish National Corpus, I show to what extent epistemicity is dependent upon grammatical person and speaker deixis. The data corpus enables us to examine such instances of epistemic coding, which are used infrequently. Thus, I pay attention to these constructions, namely when the speaker and clausal subject collapse referentially.

1 Introduction

The contribution investigates modal constructions in Polish, which express both epistemicity and evidentiality. The discussion shows that if the speaker and clausal subject collapse referentially, an epistemic meaning of a modal unit is excluded. In pursuing this aim, I provide a description of the exotic modal *musieć* 'must' by looking at its semantic properties with a particular focus on speaker deixis. After characterizing the double displacement of *musieć*, I move on to discuss the interpretations of the described modal expressions. Chapter 3 concludes the paper.

2 Modal verbs

If we take a look at languages that have a modal verb system, it may be said that modal verbs function within the grammar as an exotic phenomenon. Having considered Germanic languages, for instance, one may mention the preterite-present forms, which in turn are foreign to Slavonic languages. Considering this fact, I have adopted a universal definition of a modal verb proposed by Leiss (2009) (according to Öhlschläger 1989):

Ein Verb ist dann als Modalverb to klassifizieren, wenn es neben der Grundmodalität über eine zusätzliche epistemische Lesart verfügt. (Leiss 2009, page 6)

[A verb can be classified as a modal verb if it has at its disposal, in addition to the root modality, another epistemic reading. Translation: Ł. J.]

Roughly speaking, Polish does not have at its disposal as many modal verbs as German or Dutch. Nevertheless, we can find some Polish modal units which correspond with our definition and which can be brought down to a common denominator, i.e. to a polyfunctional one (*móc* 'can', *mieć* 'haben' and *musieć* 'must' exemplify the most used modal units). In terms of polyfunctionality, I confine myself to two main modal readings, namely deontic and epistemic. In what follows, I scrutinize the modal verb *musieć* 'must' on the basis of the Polish National Corpus. In order to go into detail, I will carefully examine constructions relating to the first person singular, both masculine and feminine forms.

2.1 The distribution of Polish *musieć*

It has been argued that the Polish modal verb *musieć* owes its existence to intensive contact with German in the 12th century. Quite in line with Hansen (2000) we do not know, however, whether Polish adopted it directly or via Czech. Nevertheless, it is worth noticing that we can find in the 14th and 15th century solely a deontic reading of *musieć*, i.e. without its polysemy, and this deontic reading is still present in the Contemporary Polish, as can be seen in [1]:

[1] *Wszystko w życiu musiałem wypracować sam.*
everything in life must [1.PS.SING.MASC.PAST.] work out
'I had to work out everything on my own in life.'

In the course of time *musieć* has obtained an additional meaning, namely an epistemic one. What we usually have in mind when we talk about this phenomenon is grammaticalization. The grammaticalization, as a kind of languages change, is a process whereby lexical items become less lexical and thus they develop or take over a grammatical function.

Epistemic modality evaluates proposition. It expresses an attitude of the speaker towards the truth of a proposition, as is presented in [2]:

[2] *Musiałem rabusiów spłoszyć podjeżdżając autem.*
must [1.PS.SING.MASC.PAST.] robbers scare away driving up car
'I must have scared the robbers away driving up in the car.'

Musiałem both in [1] and [2] function as propositional operators, in contrast to lexical verbs such as *wypracować* 'work out' or *spłoszyć* 'scare away'. They are, however, some differences between these two modals. *Musiałem* in [1] belongs to the circumstances; it governs not only the subject but also the bare infinitive, whereas the modal verb in [2] does not. *Musiałem* in [2] involves the speaker's conclusions about the circumstances. Moreover, one should not overlook the fact that the modal in [2] has wide scope, while *musiałem* in [1] has narrow.

Now, let's take a look at the data from the Polish National Corpus. All constructions consisting of the modal verb *musieć* in the past tense and the 1st person singular have been selected from the corpus and analyzed. Furthermore, the Polish language, as the majority of Slavic languages, marks genders by means of person ending. In terms ofthe research, I have found a total of 12341 tokens. Considering the gender distinction, table 1 gives the outcomes of the masculine usage (i.e. *musiał**em***), - while table 2 shows the results of the feminine usage (i.e. *musiał**am***):

deontic modals	epistemic modals	total
7748 (91,6%)	708 (8,4%)	8456

Table 1. Readings of *musiałem* in the Polish National Corpus

deontic modals	epistemic modals	total
3723 (95,9%)	162 (4,1%)	3885

Table 2. Readings of *musiałam* in the Polish National Corpus

The data shows that both interpretations occur simultaneously in the Contemporary Polish. It must be pointed out, however, that if we take a look at the statistical usage of *musiałem* and *musiałam,* the deontic reading takes precedence over the epistemic. While over 90% of tokens are attributed to the deontic modality, epistemicity seems to be overshadowed. In the next subsection, an attempt will be made to characterize the epistemic constructions.

2.2 The double displacement

There is no doubt that such epistemic constructions, in which the speaker and subject collapse, imply that the speaker refers both to himself and to his knowledge status. He assesses a situation in which he was forced to act. At this point, it is interesting to note that we come across another grammatical category in addition to epistemicity, namely evidentiality. A sentence like [2] points to a source of the assessment and the marking of the source of the information of the statement is known as evidentiality[1]. The coincidence of epistemicity and evidentiality is labaled as the double deixis (cf. Abraham 2008b) or the double displacement (cf. Leiss 2009), and is reflected in Fig.1.:

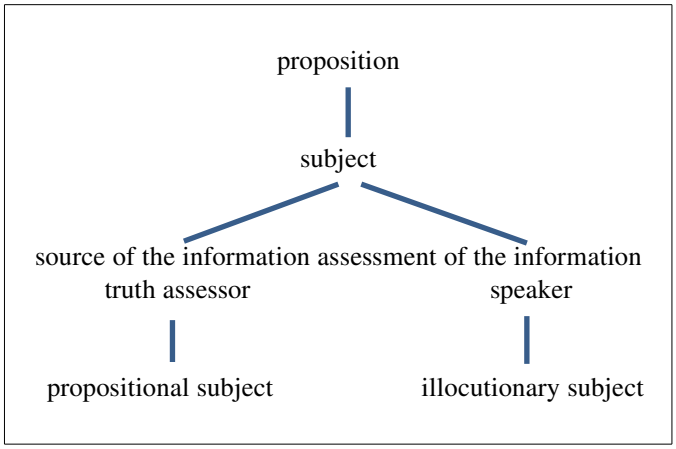


Fig. 1. The double displacement of *musiałem/musiałam*[2]

[1] It must be remarked that Polish does not have at its disposal pure grammatical markers of evidentiality, as can be observed, for instance, in Bulgarian (cf. Sauerland and Schenner (2007)).

[2] The figure has been adopted from Leiss (2009).

According to the Fig. 1., the clausal subject is divided into two parts: truth assessor, who identifies source of the information, and speaker, who evaluates the information. Taking into account this implication, it might be maintained that propositional subject and illocutionary subject are conterminous, since they refer to the 1st person. However, it is erroneous to follow this argument. In other words, the illocutionary subject distances itself from the propositional subject, because it indicates the lack of confidence in the truth of expressed proposition. This discursive function of the *I*-perspective does not pertain to the 1st person. What it means, is that the propositional subject has been shifted to the 3rd person. In this way, the speaker acts as if he were another person, whom he relies upon[3].

3 Conclusion

The weightiest result of this investigation is that as can be seen on the basis of the Polish National Corpus, the full-fledged modal auxiliary *musieć* may be characterized not only by deonticity but also by epistemicity, which in turn is bound up with evidentiality. The implication of this findings is that *musieć* may code two "grammatical" categories at the same time, on the one hand the source of information and its assessment, on the other. In both cases, we have to do with the 1st person which serves to assess proposition and therefore functions as a source of the assessment. Additionally, one may say, that at first glance, the multiple deixis rests upon the person in proposition, who might also be equivalent with the 1st grammatical person. Having analyzed the Polish data, however, it can be concluded that although the speaker refers to himself, he does not identify with the preposition subject. The speaker distances himself from himself in such a way as though he were another person.

References

[1] Abraham, W. (2008a). Aspektuelle und sprecher- bzw. personsgebundene Bestimmungskomponenten deutscher Modalverben. In: *Northern voices. Essays on Old Germanic and related topics offered to Professor Tette Hofstra*, pages 250–269, Dekker, K., MacDonald, A., Niebaum, H. (eds.). Peeters.

[2] Abraham, W. (2008b). *Illocutive force is speaker and information source concern. What type of syntax does the representation of speaker deixis require? Templates or derivational structure?* Manuscript, University of Munich.

[3] Bredel, U. (2002). "You can say *you* to yourself". Establishing perspectives with personal pronouns. In: *Perspective and Perspectivation in Discurse*, pages 167–180, Graumann, C., Kallmeyer, W. (eds.). Amsterdam/Philadelphia.

[3] It is worth mentioning that the pronoun shift is also applicable to the pronoun *you*, as it has been emphasized by Bredel (2000): "In empirical conversational situations, […], the usage of the addressing pronoun *du* (*you*) as a linguistic sign does not always refer to the hearer, but may also address the speaker" (Bredel 2002, page 167).

[4] Hansen, B. (2000). The German modal verb *müssen* and the Slavonic Languages – The Reconstruction of a Success Story. In: *Scando Slavica* 46, pages 77–93.
[5] Leiss, E. (2009). Drei Spielarten der Epistemizität, drei Spielarten der Evidentialität und drei Spielarten des Wissens. In: *Modalität. Epistemik und Evidentialität bei Modalverb, Adverb, Modalpartikel und Modus*, pages 3–24, Abraham, W., Leiss, E. (eds.). Stauffenburg.
[6] Öhlschläger, G. (1989). *Zur Syntax und Semantik von Modalverben des Deutschen*. Tübingen.
[7] Sauerland, U. & Schenner, M. (2007): Embedded Evidentials in Bulgarian. In: *Proceedings of Sinn und Bedeutung*, 11, pages 495–509, Pui-Waldmüller, E. (ed.). Pompeu Fabra.

A Russian EFL Learner Corpus from Scratch

Olga Kamshilova

Herzen State Pedagogical University of Russia
Saint-Petersburg, Russia

Abstract. Since Sylviane Granger's research group findings (1998) Learner Corpora (LC) have been looked upon as a valuable resource to pinpoint widespread and recurrent learner errors, predominantly lexical. Online text tools now available suggest working with grammatical issues, and their application, handy enough, redirects a traditional linguist from applying to national or other "ready-made" corpora to creating specialized "home-made" ones. Therefore "from scratch" in this case is, by no means, ignoring the previous experience, but developing a target-specific corpus structure. A written and spoken English corpus from Russian students of intermediate level was set as a research resource aimed at looking for other language events than those analyzed by Sylviane Granger's project, namely the range, complexity and frequency of sentence patterns in school essays and dialogues. The desiderata at present are comparative rather than didactic issues.
Keywords: special corpora, corpus format, EFL learner corpus, parsing in corpora, basic structures, patterns, "structural" poverty

1 Introduction

LC, that have been widely in use since the end of 1990s, are large electronic collections of written and, later, spoken texts produced by learners of English of different levels such as *International Corpus of Learner English (Louvain, Belgium)*, *Hong Kong Academic English*, *Michigan Corpus of Academic Spoken English*, *English as Lingua Franca in Academic Settings (Tampere)*. They have provided empirical means for two research traditions: that of second language acquisition (SLA) theory (North America) and that of EFL teaching (Europe). The strong tradition of LC analysis in Europe (S. Granger) revolutionized dictionary-making: the use of corpus data transformed not only the process but also the product, and dictionaries designed for learners of English have improved enormously (Granger, Rundell 2007). One strong point in SLA theory is *interlanguage* (or IL, after Selinker 1972) – a "transfer" grammar of a learner language that displays systematicity and deliberate choice rather than random error of other than native patterns, structures which are supposed, with the growth of literacy, to be replaced by native-like ones. IL analysis of higher-level EFL speakers, however, reveals that these preferences remain (Hinkel 2003, Mauranen 2003). Moreover, the "transfer" English grammar of speakers of different national languages tends to unification. This is actually the grammar that ensures successful ELF (English as Lingua Franca) communication.

2 Hypothesis

IL analysis even at intermediate level may reveal the most frequent patterns and structures that make up the core of EFL speakers' actual grammar. These patterns and structures are definitely native language oriented and only partly cover the list of Basic English structures.

LC data analysis seems to be a most appropriate method to test the hypothesis. To carry out the analysis there was a need for a special corpus that was launched as a Master's degree Program project (Камшилова, Колина, Николаева 2008). There were several reasons for a new corpus. *Firstly*, the access to the existing LC data is not free. *Secondly*, the available observations of learner text production are mainly about vocabulary and pragmatic issues with little concern for morpho-syntactic problems. *Thirdly,* most LC are compiled of advanced level texts, a level at which the study modes and personal objectives as well as proper instruction provide a close to native-like performance in written text production due to practiced patterns and format. *Fourthly,* the existing LC do not include (with the exception of ICLE) English texts produced by Russian students, so compiling a LC of Russian mother-tongue learners seemed both opportune and challenging. Besides, the project aimed at supporting the idea that with numerous free tools and techniques of analysis now available a special corpus is becoming an efficient instrument in language research.

3 Corpus format

The target LC (SPb LC) compilation was based upon the Lerner Corpus design criteria (Granger 1998:8) and use of free text tools now available that suggest working with grammatical issues in accordance with the research objective.

The authentic EFL textual data were collected in Saint Petersburg schools in November-December 2007. The contributors were 78 pupils of high school (forms 9–11) with pre-tested language proficiency as intermediate (26%) and upper-intermediate (74%)

The corpus size now is 38 122 words.

The language/text relevant criteria included *medium, genre, topic, technicality and task setting.* The corpus contains two subcorpora of written texts (essays and personal letters) and two subcorpora of oral monologues and dialogues (in scripts). The genres and the topics were suggested by the school syllabus and the format of the most recent innovation into the State Certification system of Russian School Education – ЕГЭ (State General Exam). Task setting was deliberately different from the requirement of S.Granger's corpus: since the informants were quite familiar with the topics, the text production was timed and a size limit for written text was set. No reference tools (dictionaries or grammars) were used either.

The learner relevant information was collected in questionnaires and later shaped the headers. The standardized header eventually included the following data:

Text Type: Essay
ID: Paul
Age: 17
Sex: female
School: 98
Form: 11
Mother-Tongue: Russian
L2: Swedish
Additional Language Training: no
Language level: Intermediate
Topic: Some people "live" in the Internet, they are real "websters", while others never navigate the Internet
Wordcount: 198 words
Date: 14.12.2007

The corpus was annotated with the help of free corpus managers (see the Reference), the tools that produce such information as word counts, frequencies, and collocation and syntactic patterns. Each text, as a result, is presented in the corpus in three modes: original, parsed and error-tagged.

The choice of the corpus managers, parsers and error-taggers was guided by the research objective (Разумова 2008). The authentic learner output data contain a high rate of non-standard forms due to many spelling and grammar mistakes. That is why some degree of caution should be exercised with standardized text retrieval software (Granger 2002:15). The small size of the SPb LC is, in this way, rather an advantage, since a manual correction can be applied.

The corpus tools include a corpus builder that is meant for the corpus completion.

4 Corpus findings: frequency list analysis

	French LC	**Quebec LC**	**SPb LC all texts**	**SPb LC Essay**	**SPb LC Letter**	**SPb LC Monologue**	**SPb LC Dialogue**
1	the	the	I	to	I	I	I
2	of	to	to	the	to	and	you
3	to	I	and	and	you	to	to
4	a	a	you	I	and	my	the
5	and	of	the	is	in	a	like
6	is	and	a	a	the	have	and
7	in	in	is	it	my	is	we
8	that	that	it	you	a	like	go
9	it	is	in	people	very	in	it
10	be	it	like	have	is	very	think

Table 1. Top 10 word forms in tree LC

A comparative frequency list analysis of three raw learner corpora – French learner writing corpus (ibid:17), Quebec advanced LC (Cobb 2003) and the SPb LC – is only a rough exploratory survey, but it provides some interesting perspectives.

It should be stressed again that the learner language proficiency varies in the corpora from advanced in the French and Quebec LC to intermediate/upper-intermediate in SPb LC. Besides, the task setting criteria were different: the contributors to SPb corpus were set a time limit and did not use any reference materials so that the text production was *nearly* spontaneous (except for the previous class practice). Therefore, both the vocabulary and the sentence patterns presumably reflect that actual language fund that the learners resort to in case of FL communication. The rich vocabulary and developed sentence patterns trained in class would give way to simple, common lexis and transparent structures. When compared with what is vaguely known as Basic Grammar of English, these structures may, hypothetically, shape the core of IL.

The top 10 in the compared corpora suggest that the task subjects in SPb corpus were definitely 1st-person oriented, hence the first rank of "I", but for the Essay subcorpus. The reason is evident: the essay format, which results in like data for "to" and "the" in the three corpora. The attraction is the high frequency rank of "and", as well as the verb forms "like", "have" and "is" in SPb LC. As the concordance display shows, the conjunction "and" is used to connect short and numerous clauses, homogeneous parts, to start a sentence and to fill the pause in case of hesitation – all that features in spontaneous speech production. The appearance of "have" and "like" in The Top 10 List, which are with a minor exception used as notional ones (the Essay subcorpus finds 109 hits for "have", with only 4 auxiliaries and 14 modals) suggests that SPb schoolchildren make wide use of the patterns I/WE <HAVE> Y and I/WE <LIKE> C. The high rate of "think" in the Dialogue subcorpus is due to a standard opinion phrase (I THINK) at the beginning (preferably) of the sentence. In the comparable corpora the only verb forms that are found on the list are the forms of "be" which can be either notional, link, modal or auxiliary. Though non-specified, this rank suggests a greater variety of verb patterns and, consequently, sentence structures in advanced learner output.

In SPb LC the verb form "is" was used 413 times. The cluster analysis demonstrates 22 notional uses (*there is, is in*), 4 modal (*is to*), and 11 auxiliary ones (3 for passive – *is called/fried* and 8 for progressive – *is working/keeping/ making etc*.) So, the major use is in the link function in the pattern X IS Y.

5 Corpus findings: basic grammar patterns

Learner preference in pattern and structure choice was studied semi-automatically in the parsed texts section of SPb LC with reference to concordance, frequency list and collocation list. Following the hypothesis, the corpus data can pinpoint the most frequent grammatical structures, used by foreign learners systematically and delibe-

rately, that shape the basic grammar of IL. The question is whether the list of preferred structures bears any resemblance to the Basic English Grammar structures and, if it does, what are the differences that bring to "structural poverty" and "simplicity without elegance" (Hinkel 2003).

Though the idea of Basic Grammar is commonplace, there is more than one attempt to its formal description (from the famous descriptive and structural models to schemes for EFL teaching). The list taken as model in this paper is adopted from Longman Grammar of Spoken and written English (1999):

- S V A (Mary is in the house);
- S V C (Mary is kind / a nurse);
- S V Od (Somebody caught the ball);
- S V Od A (I put the plate on the table);
- S V Oi Od (She gives me presents);
- S V (The child laughed).

A sample preliminary observation in the parsed corpus section provides convincing evidence of the Basic structures employment in learner text production. Each of the listed sentence models is found in learner texts. However, it is the repetition of high frequency structures and the actual fill of the structure components that imparts the non-native "accent".

It is a commonly accepted fact that the models SVC and SVA (basically with *be* as the link) prevail in non-native texts. The SPb LC data display the like tendency: the relative frequency of predicates with *be*-link is 0.33.

V-components repertoire in S V Od, S V Od A, S V Oi Od, and SV patterns demonstrates another feature of the actual learner grammar which is a deliberate avoidance of perfective and progressive forms. This avoidance is felt by the native speaker as a failure to nuance the aspect of the event/situation described, but for the IL communication, it seems to be no obstacle. The preferable modal operator is *can* (with absolute frequency in the corpus is 235 it ranks 20).

The frequency list analysis pinpointed the prevalence of *have* in essays and *like* in monologues as compared to other notional verbs. It is worth considering the high rank as the result of resorting to simple IL structures.

As the frequency list does not specify the part of speech, further analysis was carried out with reference to the parsed texts section. The parser used for this purpose was the demo version of ENGCG due to its being immune enough to the non-standard learner text and for its convenient tags. The parsed texts analysis found 362 hits for *have* in the whole SPb LC and 114 in the Essay subcorpus. Notional use of *have* was found in 323 contexts with 114 contexts in the Essay subcorpus. The modal use is high enough – 35 and 17 respectfully, while the use of auxiliary *have* (for Perfect forms) is insignificant (4) and often inaccurate:

[*I "have been falling in love" with "Counter-strike" for 6 years*][1]

The absolute prevalence of notional *have* suggests a wide use of the pattern S <HAVE> Od. The suggestion is definitely backed by the cluster display that highlights the usages of the type:

[*I also have a cat and fish*].

What seems quite special for the learner texts is the overuse of the pattern with the Od position filled with nominalized forms, that is, they prefer a precast pattern to a more conventional for native speakers nominal (*be allergic, be free, be independent*) or verbal predicate (*to communicate*):

[*I can understand when people don't keep any animals because they have an allergy*]
[*If you leave your childhood house you'll have your own life*]
[*If you have your own accomodation you also have a freedom*]
[*I enjoy to have communication with interesting people from different countryes*].

Like is a frequent character in the whole corpus (359 hits), but since it is on the Top 10 list in the Monologue subcorpus (139) the data of this part of the LC seem to be of primary interest.

The ENGCG parser marks the notional use of *like* as @+FMAINV in 117 contexts. Combinations like *I like studying* are marked as complement structures – @CS %CS (2). Twenty more word forms were marked as prepositions. The parser failed only 10 times with the non-standard learner pattern (error) of the type [*I very like my family; I very like studying English; I very like to swim there*].

The 117 contexts with the S <LIKE> C structure with Infinitive in the position of C indicate that Russian learners do not differentiate the nuances of Infinitive and *-ing* complements and systematically restore to the pattern with the Infinitive complement similar to the Russian one (*Я люблю читать*). In some monologues the pattern is numerous:

[*I'm 15 years old. I was born in Krasnoyarsk and we moved to St. Petersburg when I was 5 years. I live with my mother, father, sister and brother. My mother is a doctor, father is an architect, my brother is 25 and he is a musician, he has a group. It's very cool to have a big brother my sister is 10 and she goes to school. We have a cat and a dog. I have a lot of friends, my best friend is in my class, and we are emm we like to do something together in our free time. My friends are funny; we have fun together, make parties, and do some crazy things. We meet with our company every weekend.* ***I like skating*** *and I have a lot of friends skaters. In winter I want to learn...to learn snowboard, but there is no snow.* ***I like to draw*** *and I want to learn to draw graffiti. I'm emo,* ***I like this style*** *and a lot of my friends are emo too.* ***I like***

[1] The square brackets here and further on mark the authentic learner text with all mistakes preserved

piercing**, I have one in my tongue, I want to do one in my...right here, under my lip, but **my mother doesn't like it**. **I like rock music** and my favorite group is My chemical Romance. **I don't like pop music**. Sometimes I write my own poems and songs and stories. We often travel with my family; we sometimes go to Finland by car. **I like traveling** and I dream to visit Germany - **I like their language and culture**. **I also like cartoons and different films and cinema**. **I like coffee and candies.]

6 Conclusion

Writing and speaking with timed task setting calls for a repertoire of constructions and patterns that are simple enough and are frequently repeated in the learner output. Their simplicity is backed by their resemblance to the mother-tongue patterns. Their systemic preference to well-trained and more appropriate ways of expressing the same concepts, as intermediate and upper-intermediate levels of language proficiency demand, brings to the so-called "structural poverty". At the same time this systemic preference is the proof of learner's interlanguage grammar that, very likely (since such structures can be found in the output of non-native speakers with higher levels of proficiency), shapes the core grammar of EFL.

Such systemic preferences are best shown in LC. The corpus data supply the researcher with evidence that can be analyzed with different text tools. SPb LC is an attempt to compile a target-specific structure, a text collection in accord with essential corpus design criteria. Operated with reliable free tools, the corpus proved efficient enough in spotting and analyzing the learner language. The SPb LC findings are preliminary observations of learner interlanguage grammar.

References

Electronic

[1] AntConc [http://www.antlab.sci.waseda.ac.jp]
[2] AntConc 3.2.1 [http://www.antlab.sci.waseda.ac.jp/software.html]
[3] Connexor Natural Knowledge (ENGCG) [http://www.connexor.eu/technology/machinese/demo/syntax]
[4] The Compleat Lexical Tutor for Data Driven Language Learning on the Web [http://www.lextutor.ca/]
[5] Find and replace [http://www.abacre.com/afr.zip]

Paper

[1] *Biber D., Johansson S., Leech G., Conrad S., Finegan E.* Longman Grammar of Spoken and Written English. Harlow: Longman, 1999.

[2] *Cobb T.* Analyzing late interlanguage with learner corpora: Quebec replications of three European studies // Canadian Modern Language Review, 59(3), PP.393-423, February 2003.

[3] *Gaskell, D., Cobb, T.* Can learners use concordance feedback for writing errors? Submitted to System, November 2003, Revision April 12, 2004 [http://www.lextutor.ca/concordancers/concord_e.html]

[4] *Granger, S.* Computer learner corpora, second language acquisition and foreign language teaching. – Amsterdam: J. Benjamins Pub. Comp., 2002

[5] *Granger, S., Rundell, M.* From corpora to confidence // English Teaching Professional, Issue 50, May 2007, PP.15-18 [www.etprofessional.com]

[6] *Hinkel E.* Simplicity without elegance: Features of sentences in L1 and L2 Academic texts // TESOL Quarterly, Vol.37, No.2, 2003, pp.275-301.

[7] *Mauranen A.* The Corpus of English as *Lingua franca* in Academic Settings // TESOL Quarterly, Vol.37, No.3, 2003, pp.515-527

[8] *Selinker, L.* Interlanguage // *International Review of Applied Linguistics 10,* 1972, pp. 209-31.

[9] *Камшилова, О. Н., Колина, М. В., Николаева, Е. А.* Разработка корпуса текстов петербургских школьников: задачи и перспективы // Прикладная лингвистика в науке и образовании. – СПб., 2008. – С.92-98.

[10] *Разумова В. В.* Выбор синтаксического анализатора для анализа сложных предложений в представительном корпусе текстов // Прикладная лингвистика в науке и образовании – СПб., 2008. – СС.150-154.

Preliminary Analysis of a Slavic Parallel Corpus

Emmerich Kelih

Institut für Slawistik, University of Graz, Austria

Abstract. The focus of this paper is on a detailed description of a newly-developed parallel corpus of Slavic languages. It consists of 11 Slavic translations of the well-known Russian socialist realist novel "Kak zakaljalas' stal'/How the steel was tempered" (KZS), written by N.A. Ostrovskij in the years 1932-34. The KZS contains the Slovene, Croatian, Serbian (ekavian), Macedonian, Bulgarian, Ukrainian, Belorussian, Slovak, Czech, Polish and Upper Sorbian translations. Thus, for the first time a parallel text of almost all Slavic standard languages is available. In addition to the discussion of some text-specific issues of KZS, an explorative statistical analysis and a linguistic interpretation of text length and the Type-Token Ratio is offered.

1 Introduction

Parallel texts and parallel text corpora play a crucial role in corpus linguistics, linguistic typology and text linguistics. A parallel text is a text or part of a text placed alongside its translation in one or many[1] languages ([19, 47ff.], [17, 121ff] and [25, 73]). Parallel corpora are explored in general linguistics and language processing. With respect to Slavic languages in particular, the well-known "Multext East" project (cf. [6] and ([7]), which contains many translations of George Orwell's "1984" into Slavic languages (Bulgarian, Croatian, Czech, Resian, Russian, Serbian and Slovene) and the ambitious "Regensburg Parallel Corpus of Slavic Languages" [28], which includes different translated texts from and into Russian, Belorussian, Croatian, Serbian, Slovak, Czech and Ukrainian, have to be mentioned.

According to our knowledge, however, despite the availability of these "larger" parallel corpora focussed on Slavic languages, no parallel text corpus[2] with one original text in a large number of Slavic standard languages exists. To overcome this deficiency, a new parallel text corpus, containing in sum 11 Slavic translations of the well-known Russian socialist realist novel "Kak zakaljalas' stal'/How the steel was tempered" (hereafter KZS) has been compiled by the author for a systematic cross-linguistic quantitative analysis of Slavic languages from a synergetic and quantitative point of view ([15] [16]). In our paper only a few selected problems can be outlined. Firstly, a short overview of

[1] [5, 95] recently introduced the term "massively parallel texts", defined as a huge text corpus with translations preferably in many (genetically) diverse languages, such as, protocols of the European Parliament in over 30 languages, translations of the "Universal Declaration of Human Rights" and of the Bible into more than 100 languages, etc.

[2] For more parallel-corpora projects dealing with Slavic languages cf. [28, 123] and [8], with a project on parallel word lists of Western Slavic languages. [23] mention different available – not solely Slavic – translations of "Le Petit Prince".

linguistic applications of parallel texts is given; secondly, a description of the project in progress is presented in detail; and finally the initial results of a preliminary statistical analysis of the sample size and the Type-Token Ratio of the translations is given.

2 Parallel texts and their linguistic applications

From a linguistic point of view,[3] parallel texts are an important empirical data base. They can be used for many linguistic purposes, such as:

1. For typological and cross-linguistic analyses on the phonological, morphological, syntactical and lexical level (cf. [11], [27], [5] and [29]); in particular, parallel texts are a valuable basis for the study of quantitative features of language cf. ([2, 63-64]).
2. For the analysis of the quality and the linguistic structure of translations (cf. [1], [11], [12], [9], [20] and [24]).
3. For the examination of hypotheses from quantitative and synergetic linguistics. Parallel texts provide a deeper understanding of self-regulation mechanisms of translated texts; in particular, it is not known whether these special kinds of languages abide by language laws, such as the Zipf and the Menzerath law.
4. For comparative text linguistic analyses, including the investigation of stylistic features.
5. For the study of inter-lingual readability – a research field that thus far has hardly made use of parallel texts.

Despite this broad and appealing spectrum of various applications of parallel texts, some remarkable arguments have been raised in the past concerning the limited linguistic and methodological usefulness of parallel texts have been raised. Sometimes it is argued that translated texts lack "authenticity" and "quality" (cf. [22, 102], [29, 128]). Thus some kind of "unnaturalness" of translated texts is postulated and intensively discussed in corpus linguistics as the problem of "translationese" (cf. [18] [19, 49]). This problem is hardly to be avoided in general, but except for this legitimate objection, the outstanding attractiveness of parallel texts lies in the comparision of semantically-identical or nearly-identical texts from different languages cf. [29, 130]. Furthermore parallel texts are at least written in a grammatically and morphologically "correct" way, and hence they are, depending on the examined linguistic hypotheses, furthermore an adequate and powerful empirical data base for cross-linguistic studies.

3 "How the steel was tempered (KZS)" in twelve Slavic languages

The KZS is the empirical database of an ongoing research project on quantitative phonology, namely the investigation of interrelations between the size of phoneme inventories,

[3] The paper focuses on a description of some basic quantitative features of the KZS, and no further language processing issues are discussed. For further information about sentence alignment and POS-Tagging of parallel texts cf. [27].

phoneme frequency, distribution and syllable structure in Slavic languages. In view of the lack of available and accessible parallel texts in many Slavic languages, the Belorussian, Ukrainian, Czech, Polish, Slovak, Upper-Sorbian[4], Bulgarian, Croatian (ijekavian), Macedonian, Serbian (ekavian) and Slovenian translations of the novel "Kak zakaljalas' stal'/How the steel was tempered" (KZS) were recently collected, scanned and submitted to OCR. All text files were then manually proofread. They are now available as plain text.

The particular choice of this "highly" ideological text, however, is connected with some problematic issues, which should be discussed briefly here. Firstly, the authorship of the novel is disputed; secondly, the influence of reversions by the editors has to be discussed; and thirdly, the base of the Russian original for the translations must be specified.

However, today it is accepted that the novel was doubtlessly written by N.A. Ostrovskij. The final version of the novel was however "checked" by many editors, who "polished" the text from a stylistic, and especially ideological, viewpoint ([10, 121] and [4, 8]). In this respect – as highlighted in the detailed study by [10] – the canonical monographic issue of KZS from 1934 differs slightly from the chapters previously (1932-34) published separately in the Soviet literary journal *Molodaja Gvardija*. But the changes in the monographic issue of 1934 in comparison with the versions published earlier are primarily related to some negligible ideological details, whereas the macrostructure of the novel, e.g. the division into two parts, with nine chapters in each part, remained unchanged. As a rule, the canonical text, i.e. the monograph from 1934 and it translations into the different Slavic languages, were used for the KZS. Cf. [14, 124] with a more detailled list of the scanned translations and further bibliographical information.

Leaving aside the somewhat problematic production history, edition and translation of the KZS, the novel – the literary and aesthetic quality can be considered low – is a quite interesting mixture of different styles. According to [10, 140], for KZS a simple, linear sentence structure, a high frequency of colloquial elements and in part a "declamatory" register ('obščestvennaja reč") is characteristic. Additionally, a high proportion of oral speech, mixed with a few narrative sequences and some (typical) Soviet abbreviations and acronyms (especially in respect of the political "lexicon") are obtainable. Furthermore, the novel is characterised by an "internal heterogeneity": Throughout the novel many poems, diary entries, letters and public announcements can be found. However, the novel is not literarily "deformed" in the strictest sense, but rather a representative example of a literary prose text, written in a style typical for the socialist realist writing of the 30s, in which several sub-registers of written language in the original and translated texts can be analysed linguistically.

As already mentioned, at the macro level the KZS is divided into 18 chapters. Because of the time-consuming procedure of scanning and proofreading, only 10 chapters of 18 (nine chapters of the first part and one chapter of the second part, with a total of approximately 240 printed pages) of the KZS were processed for further analysis. A

[4] Special attention was paid to the Sorbian languages, at least in one of the standard languages. [26] note in their analysis of some Slavic translations of the novel *Harry Potter* that unfortunately no translations of this text are available for Sorbian.

more detailed explanation for this decision of a size limitation goes beyond the scope of the present paper. For our purposes, i.e. a systematic analysis of the quantitative structure (esp. phonological and syllable structure), the material seems to be sufficient and should moreover be understood as case study material.

4 Quantitative characteristics of KZS

This chapter provides some promising results[5] of the analysis of the text length (number of types and tokens) and the Type-Token Ratio (hereafter TTR) of the KZS. It will be shown that a quantitative characteristic as seemingly trivial as the sample size can already provide more detailed information about the morphological structure of the languages under examination. Furthermore it will be demonstrated that the TTR is an appropriate parameter for a typological ordering of Slavic languages.

4.1 Sample size: Number of types and tokens

An important characteristic of parallel texts is the text length, measured by the number of "words"; to be more precise in this context, by the number of tokens and types. Both linguistic entities are here defined as orthographical units of a written text, where the space has the function of a delimitation marker, i.e. all alphabetical signs between two spaces are defined as a token/type. The results of the counts of tokens and types in the KZS are summarised in Table 1 (see p. 179), where the Slavic languages are already arranged according to their genetic/areal affiliation. Fig. 1[6] represents the number of types and tokens.

The noticeable variety of the sample size[7] in the parallel texts, especially in respect of the number of tokens, can probably be explained by different morphological and syntactical characteristics of the Slavic languages. To start with the Eastern Slavic languages (Russian, Ukrainian, Belorussian), it appears that they share approximately the same text size with respect to the number of types and tokens. The difference in relation to Russian of 63 tokens (Ukrainian) and 335 tokens (Belorussian) is relatively

[5] For a comparative analysis of KZS grapheme frequencies cf. [13].

[6] Slo = Slovene, Cro = Croatian, Serb = Serbian, Bulg = Bulgarian, Mz = Macedonian, Sorb = Upper Sorbian, Rus = Russian, Ukr = Ukrainian, Belorus = Belorussian, Cz = Czech, Sk = Slovak and Pol = Polish

[7] The parallel texts have not yet been annotated and tagged. For Russian, Serbian and Slovene a sentence alignment (with the help of Duško Vitas, Belgrade) has already been performed. The author is grateful for any cooperation with specialists of tagging, annotation and lemmatisation of Slavic texts. Even if the KZS has thus far not been processed adequately, some quantitative studies can at least be performed. It can also be claimed that all texts bear approximately equal semantic information, and that all texts do not differ in terms of the original version used for the translations. On the one hand all scanning and proofreading has been done manually and on the other hand the text length of all ten chapters of each language was statistically correlated with the chapter text length of the Russian original. The interrelation of the text length in translated and original texts can be described by simple linear models, with a sufficient $R^2 > 0.98$ in all examined cases. Thus, it is likley to conclude that for all translated texts the same source was used and there are no significant stylistic modifications in the translations.

№	Language group	Language	Tokens	Types	TTR
1	South Slavic	Slovene	62655	13946	4.4927
2		Serbian	56230	13642	4.1218
3		Croatian	56424	13737	4.1074
4		Bulgarian	57174	12308	4.6453
5		Macedonian	58837	11465	5.1319
6	Eastern Slavic	Russian	49675	15053	3.3000
7		Ukrainian	49612	14645	3.3876
8		Belorussian	50010	14858	3.3659
9	Western Slavic	Czech	52180	14136	3.6913
10		Slovak	52099	14027	3.7142
11		Polish	52737	14978	3.521
12		Sorbian	58484	14574	4.0129

Table 1. Text length (types, tokens) and TTR of KZS

marginal, so it can be concluded that the Eastern Slavic languages do not show notable differences between the text lengths, either at the tokens or at the types level.

A relativley similar picture is obtained for the Western Slavic languages: The Slovak, Polish and Czech translations have a more or less similar text length (approx. 52300 tokens). This especially holds true for Slovak and Czech, with a difference of just 82 tokens. The text length of Polish is somewhat higher, for instance in relation to Czech, which has 550 tokens more. A clear outlier within the Western Slavic languages is the Sorbian translation, which has 6304 tokens more than the Czech text. This difference can perhaps be explained by morphological differences between Czech and Sorbian, such as the analytic form of the past tense, the formal expression of the reflexivity, etc. This is a hypothesis which of course must be studied in more detail in the future. However it is worth mentioning that on the types level the difference is not so for the Sorbian text, which has only 438 types more than the Czech text.

Further evidence for our claim that morphological characteristics are responsible for the differences in text length can be found in South Slavic languages. The genetically "very close" languages of Croatian (56424 tokens) and Serbian (56230 tokens) do not show any notable differences at all; similarly small differences regarding the number of types and tokens can be found for the Bulgarian and Macedonian texts. Again, just one language, namely Slovene, demonstrates a behaviour that is in some way specific regarding the number of tokens. Slovene, in comparision with other Slavic languages, shows a relativley "normal behaviour" on the types level, on the tokens level it is one of the longest texts (62655). This striking feature – at least on the tokens level in relation to the Russian text (49675 tokens) – can again be explained linguistically: In Slovene, the analytical form of the past tense by means of the auxiliary verb "to be/je", the intensive usage of the analytical past perfect ("je bil vedel/be was known") and the frequent use of relative clauses are relatively typical and commonly used. Russian neither makes use of auxiliary verbs nor has a complex analytical past tense form and thus for the Slovene text a frequent use of synsemantic words and a high number of tokens can be observed.

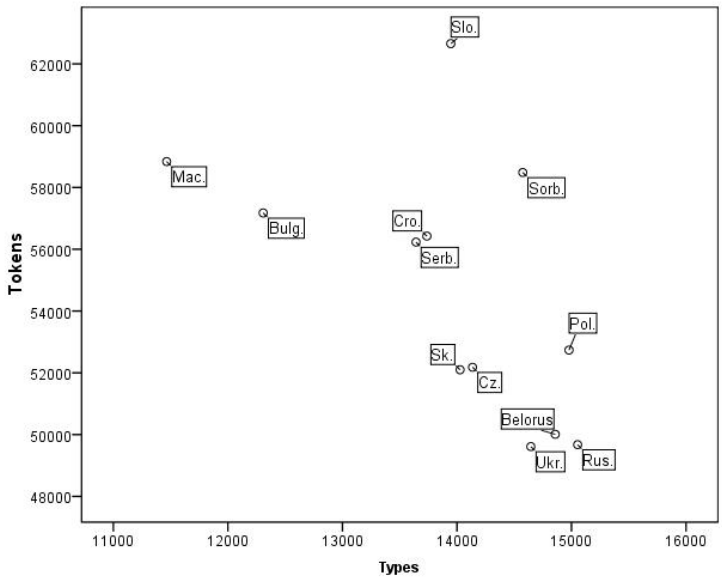


Fig. 1. Types and tokens in Slavic parallel texts

All in all, a fairly simple analysis of the text length already provides some in-depth information about the morphological and syntactical structure of the parallel text corpus under examination.

4.2 Type-Token Ratio

In addition to a simple description and interpretation of the number of types and tokens of the parallel texts, a more detailed examination of the so-called Type-Token Ratio (TTR = Tokens/Types) is required. Whereas in the past this index was primarily understood as an indicator of lexical richness (cf. the overview on possible interpretations of the TTR in [3, 108]), the TTR – especially in comparative studies of parallel texts – can be introduced as a measurement of the morphological richness of word forms and the productivity of the flexion system. The more word form types a text (=language) has, the greater is the variety of morphological forms that it contains and thus a low TTR in parallel-text research can be interpreted as an indicator of the synthetism of one language. For more in-depth research of word frequencies in language typology cf. [21]. The TTR for all examined Slavic languages is graphically represented in Fig. 2 and already ranked from its maximum to its minimum.

It can clearly be seen from Fig. 2. that, based on the TTR, a typological order of the Slavic languages can be obtained: It starts with the strongly analytic Southeast Slavic languages, Macedonian and Bulgarian, and continues with Slovene, Serbian and Croatian. They are followed by the Western Slavic languages (order of languages: Sor-

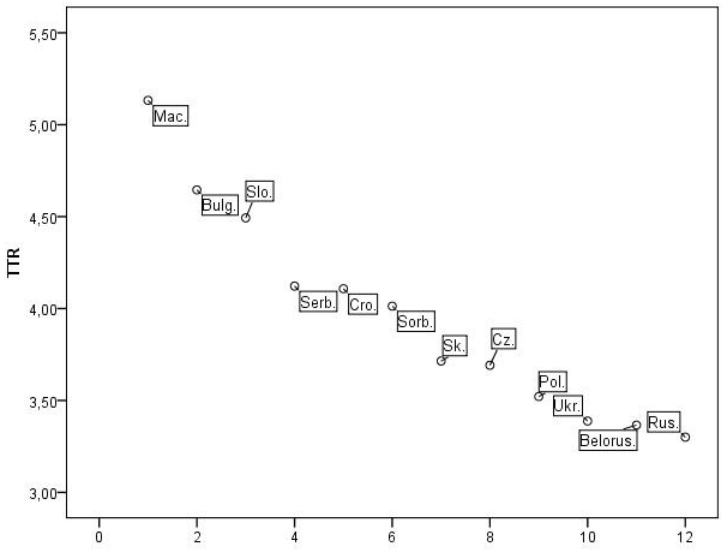


Fig. 2. TTR in Slavic parallel texts

bian, Slovak, Czech and Polish), which appear as one "group"[8]. Finally, the East Slavic languages (Ukrainian, Belorussian and Russian) have the lowest TTR and also appear as one typological "row". In sum, the TTR of parallel texts can be interpreted alongside the morphological richness as an indicator of the degree of analytism/synthetism of the languages.

5 Conclusion

The paper has the main function of presenting the KZS parallel-text corpus, which should, of course, be refined in the future, especially from the viewpoint of language processing (tagging, alignment, and lemmatisation). As shown in our exploratory discussion, the text size (number of types and tokens) and the Type-Token Ratio of parallel texts appear to be a simple yet efficient tool for obtaining language-specific morphological behaviour of the parallel texts analysed. Moreover, strong evidence for the general usefulness of parallel texts for language typology is given.

References

[1] Altenberg, B. and Aijmer, K. (2000). The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. In Ch. Mair and Chr.

[8] In future more rigorous methods such as cluster methods must be applied.

Hundt, editor, *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*, pages 15–33, Amsterdam. Rodopi.

[2] Altmann, G. and Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. Fink, München.

[3] Altmann, V. and Altmann, G. (2008). *Anleitungen zu quantitativen Textanalysen. Methoden und Anwendungen*. RAM-Verlag, Lüdenscheid.

[4] Anninskij, L. (1989). Obručennyj s ideej. In Ostrovskij, N. A., editor, *Sobranie sočinenij v trech tomach. Tom 1. Kak zakaljalas' stal'*, pages 7–28, Moskva. Molodaja Gvardija.

[5] Cysouw, M. and Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *Sprachypologie und Universalienforschung*, 60(2):95–99.

[6] Dimitrova, L., Ide, N., Petkevič, V., Erjavec, T., and Tufiș, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons and Lexicons for six Central and Eastern European Languages. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics. Volume 1*, pages 315–319, Montreal and Quebec.

[7] Erjavec, T., Ide, N., Petkevič, V., and Véronis, E. (1995). Multext-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. In *Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995)*, pages 88–97.

[8] Garabík, R. and et al. (2007). A Cross-linguistic Database of Children's Printed Words in Three Slavic Languages. In Levická, J. and Garabík, R., editors, *Slovko 2007. Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007*, pages 51–64, Bratislava. Tribun.

[9] Gellerstamm, M. (1996). Translations as a source for cross-linguistic studies. In Aijmer, K., Altenberg, B., and Johansson, S., editors, *Language in Contrast: Papers from a symposium on Text based Cross-linguistic studies. Lund, March 1995*, pages 53–62, Lund. Lund University Press.

[10] Guski, A. (1981). N. Ostrovskij: Kak zakaljalas' stal': biographisches Dokument oder sozial-realistisches Romanepos? *Zeitschrift für slavische Philologie*, 42:116–145.

[11] Johansson, S. (1998). On the role of corpora in cross-linguistic research. In Johansson, S. and Oksefell, S., editors, *Corpora and Cross-Linguistics Research*, pages 3–24, Amsterdam. Rodopi.

[12] Johansson, S. (2003). Reflections on Corpora and their Uses in Cross-linguistic research. In Zanettin, F., Bernardini, S., and Stewart, D., editors, *Corpora in translator education*, pages 135–144, Manchester. St. Jerome Publisher.

[13] Kelih, E. (2009a). Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle. *Glottometrics*, 18:53–96.

[14] Kelih, E. (2009b). Slawisches Parellel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In E. Kelih, E., Levickij, V. V., and Altmann, G., editors, *Methods in text analysis*, pages 106–124, Černivci. Ruta.

[15] Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Brockmeyer, Bochum.

[16] Köhler, R. (2005). Synergetic linguistics. In Köhler, R., Altmann, G., and Piotrowski, R. G., editors, *Quantitative Linguistik/Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*, pages 760–774, Berlin. de Gruyter.

[17] Lemnitzer, L. and Zinsmeister, H. (2006). *Korpuslinguistik: eine Einführung*. Narr, Tübingen.

[18] Mauranen, A. (2002). Will 'translationese' ruin a contrastive study? *Languages in Contrast*, 2:161–186.

[19] McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge, London.

[20] Mohanty, P. (2008). The Semantic Differential Technique and Measurement of Translational Meaning. In Altmann, G., Zadorozhna, I., and Matskulyak, J., editors, *Problems of General, Germanic and Slavic Linguistics. Papers for the 70th anniversary of Professor V.V. Levickij.*, pages 215–225, Černivtsi. Knichi XXI.

[21] Popescu, I.-I. and Altmann, G. (2008). Hapax Legomena and Language Typology. *Journal of Quantitative Linguistics*, 15(4):370–378.

[22] Stolz, T. (2007). Harry Potter meets Le petit prince: On the usefulness of parallel corpora in crosslinguistic investigations. *Sprachtypologie und Universalienforschung*, 60(2):100–117.

[23] Stolz, T., Stroh, C., and Urdze, A. (2007). Nicht ganz ohne [...]. In Grzybek, P. and Köhler, R., editors, *Exact Methods in the Study of Language and Text. Dedicated to Professor Gabriel Altmann on the Occasion of His 75th Birthday.*, pages 633–646, Berlin and New York. de Gruyter.

[24] Teubert, W. (2002). The role of parallel corpora in translation and multilingual lexicography. In Altenberg, B. and Granger, S., editors, *Lexis in Contrast*, pages 189–214, Amsterdam. Benjamins.

[25] Teubert, W. and Čermáková, A. (2007). *Corpus linguistics: A short introduction*. Continuum, London.

[26] van der Auwera, J., Schallea, E., and Nuyts, J. (2005). Epistemic possibility in a Slavonic parallel corpus – a pilot study. In Hansen, B. and Karlík, P., editors, *Modality in Slavonic languages. New perspectives*, pages 201–217, München. Sagner.

[27] Véronis, J. (2000). From the Rosetta Stone to the Information Society: A Survey of Parallel Text Processing. In J. Véronis, editor, *Parallel Text Processing. Alignment and Use of Translation Corpora*, pages 1–25, Dordrecht. Kluwer.

[28] von Waldenfels, R. (2006). Compiling a Parallel Corpus of Slavic Languages. Text strategies, Tools and the Question of Lemmatization in Alignment. In B. Brehmer and V. Zhdanova and R. Zimny, editor, *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*, pages 123–138, München. Sagner.

[29] Wälchli, B. (2007). Advantages and disadvantages of using parallel texts in typological investigations. *Sprachypologie und Universalienforschung*, 60(2):118–134.

Operators for Extending and Developing an Utterance (Based on Operators of Concessive Relation)

Jana Kesselová

University of Prešov in Prešov, Slovakia

Abstract. The paper deals with operators of concessive relation which we consider to be a transitional type between operators used for developing an utterance and operators used for extending an utterance. An empirical source of the research is the Slovak National Corpus. The aim of this research is to enrich the existing paradigmatic approach to classes of operators by a syntagmatic approach focused on their real textual "behaviour". We proceed from the form of an operator through its contextual setting towards more general aspects of its functioning. At the same time we concentrate on the inventory of concessive operators in contemporary written Slovak, types of relations between utterance meanings, and communication functions of utterances with individual types of operators. Contemporary written Slovak has an extraordinarily miscellaneous and formally, stylistically, distributionally and pragmatically differentiated inventory of concessive operators. Dynamics in this microsystem is demonstrated by strengthening the position of a prototype representative and identifying the centre with synchronically most frequent operators on the one hand and decline (even disappearance) of some synonymous operators on the other hand. Another example of dynamics is stylistic and pragmatic specification of operators.

1 Introduction

While basic word classes have been described and analysed in Slovak linguistics at least in one monograph each, non-basic word classes do not belong to the centre of linguists' research attention (one of exceptions is the period of preparation of a book *Morfológia slovenského jazyka*, 1966). Nevertheless, when forming an utterance and reaching a communicative intention, means of joining isolated naming units into coherent textual sequences are as important as the names of independent objects and attributes.

Relations between phenomena present a universal system of connections and associations or, more precisely, people interpret them as a system of mutual connections and associations and they process them linguistically in communication [5]. Among means oriented functionally towards establishing a system of relations, the article focuses on connecting expressions and their functioning in a real language interaction. The centre of connecting expressions consists of words with a primarily connecting function, i.e. conjunctions and their combinations. Other connecting means are e.g. interrogative pronouns (their connecting function is only secondary), combinations of conjunctions and pronouns, and fixed phrases consisting of conjunctions and other word classes. Thus instead of using the term conjunctions I prefer the term with a much broader sense – operators. Operators for extending and developing an utterance are understood as language means with a

connecting function which join clause elements and clauses into higher utterance and textual units. They are formally miscellaneous and their centre consists of words with a primarily connecting function, i.e. conjunctions. These are supplemented by further connecting means:

• relatives (interrogative pronouns in which the connecting function is realized only as secondary);

• co-relatives (co-relative pairs of conjunctions and demonstrative pronouns or interrogative and demonstrative pronouns);

• conjunctional expressions (combinations of conjunctions and other word classes, mainly particles, primary and secondary prepositions and adverbs).

The second reason why I prefer using the term operators to the term conjunctions is motivated by a communicative-functional approach to the role of connecting means in human communication. From the communicational point of view, operators do not have only a connecting function, but they take part in creation or modification of the communication function of the utterance which is understood as a purpose (intention) of communication. Thus I concentrate on operators not only as means of establishing links and relations between utterance meanings, but also as indicators and constituents of communication functions of utterances.

Distribution of operators used for extending and developing an utterance is based on the fact that the "lower" units enter various relations in higher utterance units and they more or less lose their autonomy. According to the relation which operators establish between utterance meanings, two types of operators may be distinguished [3]:

• Related utterance meanings preserve their high degree of autonomy and the relation of relative semantic independence. Operators present one utterance meaning in relation to another one as simply coordinated, possible, contradictory or more important. Operators in utterances with semantically independent utterance meanings are considered to be operators for extending the utterance.

• When creating an utterance, utterance meanings may lose their autonomy because they enter the relation of semantic dependence. It happens in the case when one utterance meaning is determined by another one (e.g. if some action/state is presented as a cause, purpose, condition, temporal circumstance, etc.). Operators in utterances with semantically dependent utterance meanings are considered to be operators for developing the utterance.

The study focuses on operators of concessive relation which I consider to be a transitional type between the above-mentioned types.

2 Method of research

An empirical source of the research is a large number of written texts in the Slovak National Corpus. The research is based on a so-called balanced corpus, in which publicistic texts (33,3%; doc.type=inf), scientific texts (33,3%; doc.type=prf) and belletristic texts (33,3%; doc.type=img) are present. I have used data about the total frequency of a phenomenon in the Corpus which is marked as its synchronic

communication loading in a particular corpus [17]; about the stylistic value of an operator (concordances → statistics → attribute doc. type → frequency distributions); and about the most frequent collocations (concordances → statistics → attribute: word/lema → most frequent collocations). To avoid including homonymous expressions into the total number of occurrences of individual operators, I have separated operators from other word classes with a command, e.g. [(word="hoci") & (tag="O.*")] or, where necessary, also with a negative filter or manually.

Thanks to the existence of the Corpus, it is possible to complete, specify and refine the existing conclusions (based and verified mainly on belletristic texts) and enrich the existing paradigmatic approach to classes of operators with the syntagmatic approach aimed at their real textual "behaviour". I have proceeded from the form of operator through its contextual setting towards more general aspects of its functioning. At the same time I have concentrated on:

• the inventory of concessive operators in contemporary written Slovak

The system of operators provides a set of possibilities, an offer for a producer of the text. I am interested in a) motivation of selection of a particular operator and b) competing and complementary relations between central and peripheral operators.

• types of links and associations between utterance meanings

Meaning associations and links which are created by operators have been investigated in a selective set of utterances with a given operator. With regard to a high frequency of operators, it is beyond possibilities of an individual to investigate all the utterances. It is known that one of methodologic questions of corpus linguistics is representativeness of a sample created from the "profusion" of occurrences which the corpus offers. When creating a representative set, I have taken into consideration two criteria. The criterion of contextual surroundings means that utterances with the most frequent collocations of the operator have been included in the set (surprising semantic homogeneity of expressions occurring in the surroundings of operators has been proved). The second criterion is a stratified way of utterance selection into the research sample. It means that utterances from publicistic, scientific and belletristic texts have been included in the sample equally (if the calculation of frequency distribution has shown stylistic neutrality of the operator) or unequally, according to the calculated value of operator frequency distribution.

• communication functions of utterances with individual types of operators

In complex books on Slovak morphology and syntax [9, 10] conjunctions are characterised either as a set of means serving for expressing a certain relation, or as enumeration of relations corresponding with individual conjunctions. The aim of this corpus research is to show that operators are not only means used for establishing relations, but that they together with lexical surroundings, functional sentence perspective, etc. participate in creating communication functions of the utterance.

3 Concessive relation and its character

A concessive relation is traditionally characterised within the frame of dependent adverbial relations in the Slovak linguistics [10, 11] with the explanation that dependent adverbial clauses of concession express circumstances unfavourable to the action of the main (independent) clause.

On the one hand, I consider concessive relations to be a part of causal relations in a broad sense of word. Slovník slovenského jazyka III [14] defines concession as a real or possible cause whose expected consequence (effect) in a given instance does not come. The basis of the concessive relation is to realise the causality between two phenomena which is regularly or usually realized (*x* is always/usually the cause of *y*). The concessive relation is a specific aspect of causal relation in that sense, that it includes instances which go beyond regularity or commonness. In the relation cause – consequence it "counts" with the influence of unexpected circumstances, e.g.:

(1) ***Hoci*** *má Irak ohromné zásoby ropy, nemá benzín.*

The empirical context (knowledge that petroleum is used for petrol production) leads to the expectation that petroleum resources in Iraq mean also sufficiency of petrol reserves. However, the utterance with the concessive operator *hoci* brings unpresupposed (on the basis of the presupposition *"who has petroleum has also petrol"*) even surprising conclusion (favourable circumstance *x* brings unfavourable situation *y*). In other words, operators of concessive relations construct utterances which demonstrate linguistically the fact that the relation between cause and its consequence has only relative validity. P. Karlík [5] defines a concessive relation as a relation between ineffective cause and unexpected consequence.

On the other hand, the concessive relation is very close to a contrast relation since its utterance meanings are in the mutual relation of disagreement even contrast. This "semantic nearness" is reflected either in the use of combinations of concessive and contrast operators (*hoci – ale*, *hoci – no, hoci – predsa, hoci – jednako*)or in a possible mutual substitution of concessive and contrast operators:

(2a) ***Hoc*** *si moji exspolužiaci vybavujú hypotéky a kupujú byty, mne stále vyhovuje bývanie v prenájme.*
(2b) *Moji exspolužiaci si vybavujú hypotéky a kupujú byty,* ***ale*** *mne stále vyhovuje bývanie v prenájme.*

However, it is necessary to add that the potential substitution of operators relates to non-contextual utterances; substitution of operators in the context modifies the communication function of the utterance. As for their function, concessive operators are very close to causal and contrast operators. Thus I consider them to be a transitional type between operators for developing and extending the utterance.

4 Concessive operators

4.1 Inventory of operators and their position in the utterance

The inventory of concessive operators is a very diverse system which is formally, stylistically and pragmatically differentiated. From the formal point of view it consists of the conjunctions ***hoci*** (and its variant ***hoc***) and ***síce***, and the conjunctional expressions which contain basic subordinating conjunctions and relatives:

- ***keď*** (*aj keď, i keď, ani keď, keďešte*);
- ***keby*** (*aj keby, keby aj, i keby*);
- ***aby*** (*namiesto toho, aby*);
- ***že*** (*napriek tomu, že; navzdory tomu, že; bez ohľadu na to, že; nehľadiac na to, že; nezávisle na tom/od toho, že*);
- ***čo*** (*a čo aj, a čo i, čo hneď, čo ako, čo priam*);
- ***ako*** (*akokoľvek*).

Each of them may enter a binary conjunctional expression, e.g.: *hoci – predsa, hoci – jednako, i keď – aj tak*, etc. Moreover, all operators have both, initial and non-initial positions in the utterance. The non-initial position strongly prevails in communication(the only exception is the conjunctional expression *nehľadiac na to, že*, which occurs equally in both positions). Thus the order of clauses is not motivated only by the principle of constructional iconism (the order cause – consequence in reality), but also by the communication intention of a speaker/writer (which part of the utterance meaning he/she wants to emphasize – either ineffective cause or unexpected consequence), e.g.:

(3a) *Ružomberská trénerka môže byť so začiatkom série spokojná,,* ***hoci*** *k niektorým fázam zápasov mala určité výhrady.*
(3b) ***Hoci*** *dosiaľ obce návrhy nemuseli zverejňovať, viaceré to už robili aj predtým.*

A different position of the operator in the utteranceis a stylistic indicator (for example, the non-initial position of the operators *hoci, aj keď, i keď* is typical of belles-lettres style while the initial position is typical of publicistic texts). The Corpus enables us to find out even more minute differences. Different positions of operators in utterances are bound with different collocations which leads to modifications in communicative functions of utterances. The operators *i keby, aj keby,* when being in the initial position, co-create utterances in which the speaker admits a very improbable, even unreal circumstance in favour of ensuring the addressee about the truth value of his/her claim (***Aj keby som mal*** *skončiť na šibenici...*, ***Aj keby som mal*** *zomrieť dvakrát...*, ***Aj keby som mal*** *sto rokov sedieť v base...*, ***Aj keby som mal*** *do konca života kopať kanály...*, ***I keby som mal*** *k dispozícii tisíc ľudí...*, ***I keby som mal*** *sto rúk....*).

(4a) ***I keby som mal*** *donútiť rieky, aby tiekli proti prúdu, a slnko, aby nesvietilo, aj tak splním tvoje želanie*.

On the other hand, the operators *i keby, aj keby* in the non-initial position participate in utterances whose communication function is an urgent appeal to take action even in the case of unfavourable consequences.

(4b) *Ochraňujte vaše kone, zastavte ich vývoz,* ***aj keby to malo*** *byť na úkor vašich platov.*

The most frequent collocations show that unfavourable circumstances are described in an impersonal way, e.g.: ***aj keby to malo*** *znamenať vojnu*, ***aj keby to malo*** *znamenať sm*rť, ***aj keby to*** *malo znamenať zvýšenie daní*, ***aj keby to malo*** *ísť dolu vodou*, ***aj keby to malo*** *vyznieť v náš neprospech*, ***aj keby to malo*** *trvať sto rokov*, ***aj keby to malo*** *mať podobu horšieho*, ***aj keby to malo*** *znamenať, že sa stratí čas,* etc. Impersonal constructions preserve the principle of utterance quality, but at the same time they express unfavourable consequences more softly than directly addressed constructions, compare e.g.: *aj keby si mal zomrieť – aj keby to malo znamenať smrť.*

4.2 Synchronic communication loading and the stylistic value of concessive operators

Total occurrence of an operator in the Corpus reflects its synchronic communication loading. Frequency of various operators differs remarkably in communication (see diagram 1). According to J. Oravec [10] concessive conjunctions consist mainly of two bases – *hoci, čo*. The corpus research has shown that the conjunction *hoci* and operators with the component *čo* represent two opposite poles of frequency loading of concessive operators: while the conjunction *hoci* is dominant among concessive operators and it represents nearly 40 % of their uses, the operators with the component *čo* are on the decline (and are felt as literary).

(5a) *S úradmi sa nepri, syn môj, hovorievali mu otec,* ***a čo aj*** *neboli najlepším rodičom, veru mali pravdu*.
(5b) *Tento zámok nedobyjeme,* ***čo hneď*** *pod ním vycedíme ostatnú kvapku krvi.*
(5c) *Báseň je apoteózou slobody: „...****a čo i*** *tam dušu dáš v tom boji divokom...*
(5d) *Nuž* ***čo priam*** *Ľudmilôčka moja nič nedostane, veď si ju ja vyživím.*

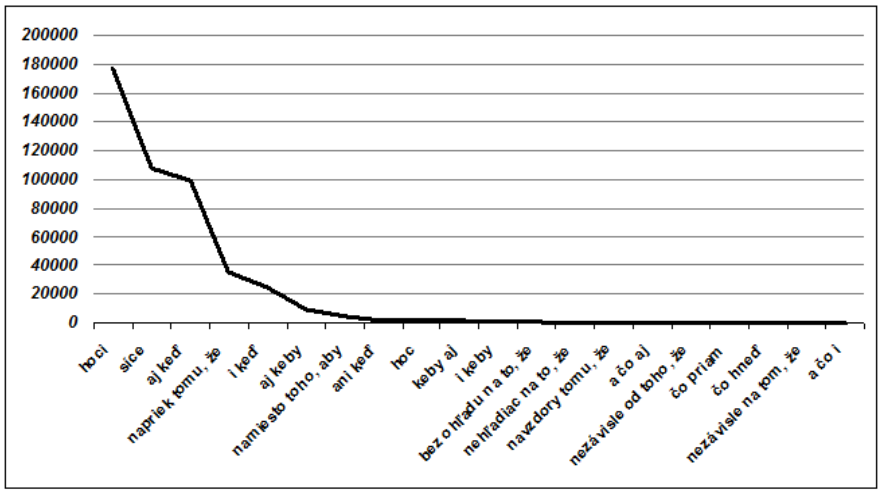


Diagram 1. Synchronic frequency loading of concessive operators

The four most frequent operators – *hoci, síce, aj keď, napriek tomu, že* cover 90 % of concessive operator uses. The operators *hoci, aj keď* are used mainly in belletristic texts, the operators *hoci, síce, aj keď, napriek tomu, že*; *napriek tomu, že* are used mostly in publicistic texts (the operators used in the initial position have a capital letter at the beginning). Similarly, other, less frequent concessive operators are used mostly in the belletristic or publicistic style; the only conjunction *hoci* prevails mainly in the scientific style.

In my opinion, stylistic distribution of operators is the outer reflection of the character of concessive relation (part 3). It is known that a scientific text is based mostly on the explanatory composition style whose basic feature is explicativeness. According to J. Dolník and E. Bajzíková [1] explication relates to statements about facts which are given; where there is no doubt, but question leading towards understanding their existence. On the other hand, concessive operators establish a relationship between contradictory utterance meanings and form contradictory statements and unexpected or surprising associations and connections. This raises doubts about presented facts and evokes the expectation that the facts are to be justified properly. Compare:

(6a) ***Keďže*** *nič neje, nepriberá.*
(6b) ***Hoci*** *nič neje, priberá.*

In the first utterance (6a) the causal operator establishes a relationship between cause and consequence in the form of a statement; in the second utterance (6b) the concessive operator introduces a contradictory statement which raises doubts and also expectation that untypical consequence will be explained or justified. The relation between causal and concessive operators resembles the relation between explicationand argumentation to the intent that causal operators are connected with the explicatory expansion of a topic and concessive operators are more connected

with the argumentation expansion of the topic. Argumentation can be either analytic or practical. According to J. Dolník and E. Bajzíková [1] practical (non-analytic) argumentation is used in common (non-scientific) communication and analytic argumentation is used in scientific texts.

Utterances with concessive operators in the analytic argumentation have a form which meets the requirements for the quality of utterance understood to the intent that we should not say anything without having sufficient evidence [15]. Concessive operators establish relationships between utterance meanings not only to emphasize various points of view, but they even stimulate recipient's multi-aspectual point of view and thus they raise possible objections, look for evidence,and finally find the truth. Compare:

(7a) ***Hoci*** *Šimonovičove preklady majú svoje prednosti, medzi ktorými vyzdvihujem najmä priekopnícku prácu s asonanciou, resp. s eufonicky nepresným rýmom, a* ***hoci*** *jeho preklady radu básní všestranne obstoja v silnej konkurencii a možno ich pokladať za adekvátne,* ***predsa len*** *funkciu základného prekladu Lorcovej poézie ako relatívneho celku aj pre slovenského čitateľa plnia skôr...)*
(7b) ***Keby aj*** *Oldřich Kulhánek, absolvent vysokej školy umeleckopriemyselnej u profesora Svolinského, nebol autorom celého emisného programu platidiel Českej národnej banky, jeho grafické dielo pre svoje kresebné majstrovstvo* ***už i tak*** *vošlo do pokladnice českej výtvarnej histórie, rozsahom námetov i hĺbkou filozofickej výpovede.*

Concessive operators also constitute utterances of practical argumentation which are not about searching for the truth in a scientific sense of the word. According to J. Dolník and E. Bajzíková [1] the practical argumentation is used to justify propositions which are beyond proving in a strict logical sense. It is not about proving the truth value of the proposition, but about showing its validity. It can be used for justifying certain behaviour or attitudes in everyday life.

8a) *Uč sa pozorovať a chápať jej situáciu,* ***hoci*** *má chyby, predsa môžeš čerpať z jej skúseností.*
(8b) *Aj rodičia boli donútení k tomu, aby dali výpoveď,* ***hoci*** *inú robotu ešte nemajú.*
(8c) *No* ***hoci*** *do Škandinávie odcestoval,* ***predsa*** *tam nezostal a po poslednej previerke vo Švédsku sa vrátil domov.*
(8d) *Obedy mu prihrieva stará pani zo susedného bytu, prilipol k nej,* ***hoci*** *s cudzími sa zbližuje len ťažko.*

In both types of argumentation it is possible to evaluate the communication function of utterances with concessive operators as justifying or giving reasons (an effort to show or prove validity of a proposition) or as ensuring the addressee about the truth value of the content which seems to be improbable, unpleasant or contrary to the existing experience. In my opinion, concessive operators belong to traditional ensuring means such as, for example, explicit performative formulae, e.g. *ubezpečujem vás* or modality particles.

4.3 Communication functions of concessive operators

Preference of some of competing concessive operators is motivated by a communication intention of a speaker. Utterances with concessive operators contain a statement and also correction of a conclusion which is made by a recipient (on the basis of the given statement). Thus the utterance meaning joined by the concessive operator restricts, relativizes or more or less denies the truth value of the proposition.

(9a) *Podarilo sa im urobiť vynikajúci album plný skvelých piesní,* ***aj keď*** *nie sú celkom originálne.*
(9b) *Formálne rozhovory na túto tému neprebehli,* ***hoci*** *priznáva, že sa o tom v kuloároch rozprávalo.*
(9c) *Doteraz to bolo celkom idylické,* ***i keď*** *nie celkom neproblematické.*

With regard to a large number of examples found in the Corpus (the number of utterances with concessive operators presented in this paper is more than 460 thousand in the Corpus prim-4.0-public-all), I have worked with selected sets created according to two criteria (most frequent collocations and stratified selection based on stylistic characteristics). Results can be characterised as tendencies which reveal a large number of real syntagmatic realizations but they definitely "resist" any rigorous classification criteria. Beneath I present some of the results.

4.3.1 Right-hand collocations of the four most frequent operators (*hoci, síce, aj keď, napriek tomu, že*) show semantic preference and semantic proximity of expressions with which operators combine in utterances. The most frequent collocations of dominant concessive operators refer to a variance between:

• how phenomena and events appear/are presented and how they really are (according to the speaker), e.g.: *navonok*, *naoko*, *zdanlivo*, *zdať sa*, *oficiálne*, *oficiálny*, *formálne*, *deklarovať*, *proklamovať*

(10a) *Bol odjakživa mojím skrytým nepriateľom,* ***hoci navonok*** *sa mi zaliečal viac, ako by človek čakal.*

• prevailing or generally accepted opinion and speaker's opinion (*väčšina, mnohí, viacerí, všeobecne, niektorí, podaktorí, väčšinovo*)

(10b) *Myslím si, že sa nám podarilo udržať určitú vážnosť architektúry v spoločnosti,* ***hoci väčšina*** *architektov to stále vidí tak, že nie.*

• situation or conditions in the past and at present (*pôvodne*, *spočiatku*, *vlani*, *pôvodne*, *predtým*)

(10c) *Nelamentuje, ani keď z nej ťahám peniaze,* ***hoci pôvodne*** *predpokladaná suma na stužkovú sa zdvojnásobila.*

A common factor of such constituted utterances is their non-operative function which oscillates between statement, evaluation and critique. O. Müllerová [8] characterizes operativeness as intensity of speaker´s influence on his/her communication partner. The aim of operational utterances is to influence partner´s behaviour; in non-operational utterances the speaker focuses more on expressing himself/herself.

Except for common tendencies, it is possible to observe also specific textual "behaviour" in syntagmatic relations. Typical surroundings of the conjunction *hoci* are verbs like *uznávať*, *priznávať*, *pripúšťať*, *netvrdiť*, *nevylučovať*, *nepochybovať*, *nepopierať*. The most frequent collocations of the conjunction *aj keď* are except for the above-mentioned verbs also quantifiers, e.g.: *aj keď* + *nie vždy*, *nie všetok*, *nie celkom*, *nie veľmi*, *nie úplne*, *nie príliš*, *nie každý*, *nie všade*, *nie celý*. The result of these combinations are utterances in which the speaker presents his/her opinion, but at the same time he/she"leaves the back door open":

(11a) *Už dávno sa nepovažuje za ľavičiara,* ***hoci nepopiera****, že vyšiel z tohto prostredia.*
(11b) *Naozaj sme hrali postupne o čosi lepšie,* ***hoci netvrdím****, že to bola prevratná zmena.*
(11c) *Šancu máme,* ***aj keď netvrdím****, že veľkú.*
(11d) *Tanečníci majú potenciál,* ***aj keď nie všetci*** *ho využijú.*

A similar communicative function, i.e. to express opinion but in a relativising way may be observed in utterances with the operators *síce* and *napriek tomu, že* in combinations with the most frequent collocations. Here belong:

• positive evaluative expressions (***síce*** + *fajn*, *chvályhodný*, *lákavý*, *efektný*, *snažlivý*; ***napriek tomu, že*** + *obsiahly, skutočný, veľmi účinný, atraktívny, dlhodobý, náročný, jednoduchý, najväčší, výhodný*);
• verbs in a negative form (***síce*** + *neuviesť*, *nespochybňovať*, *neodstrániť*, *nenahradiť*, *nezakazovať*, *nevyznačovať sa*, *nevyhlasovať*, *nepredpisovať*, *neudeľovať*; ***napriek tomu, že*** + *nesprávať sa, nehrať, nepoznať, neexistovať, nezískať*, ***síce*** + *nevynikať*, *neobhájiť*, *neskórovať*, *neumiestniť sa*, *neoslniť*, *nebodovať*, *nezvíťaziť*/*prehrať*, *nepredviesť*).

These combinations of expressions create utterances in which the speaker admits quality of a certain phenomenon but relativizes it immediately by presenting this phenomenon from a different angle. Such types of utterances may express multi-aspectual, non-simplified or untraditional depiction of things and events (12a), critique (12b) or defence reaction to critique. The speaker admits certain negatives, but immediately counterbalances them with positives (12c, d). The operators *síce* and *napriek tomu, že* occur mostly in publicistic texts.

(12a) *Snaha dostať ženy do politiky je* ***síce chvályhodná****, riešiť to však prostredníctvom zákona nie je najšťastnejšie.*
(12b) *Príslušník mladšej generácie sa v antológii nenájde, a to* ***aj napriek tomu, že*** *ide o výber naozaj obsiahly, takmer štyristostranový.*
(12c) *Kňazovický* ***síce neobhájil*** *zlato na neolympijských 200 m, ale ani bronz nie je neúspechom.*
(12d) *Zmeny v zákone* ***síce neodstránili*** *všetky nedostatky, ale určite eliminovali tie najväčšie.*

4.3.2 The producer of the text often takes an emotionally evaluative attitude to an unexpected consequence. The unexpected consequence may cause a pleasant surprise, astonishment and positive attitude. This is usually expressed explicitly (13a, b) or it follows from the utterance content. A specific place belongs to the conjunctional expression *keď ešte* which constitutes utterances with the communication function of surprise in sports and economic publicistics (13c, d).

(13a) ***Hoci*** *koncertné albumy sa zvyčajne vydávajú po niekoľkých radových platniach, Portishead sa k tomu odhodlali už po dvoch. Bol som tým veľmi milo prekvapený...* (13b) *Som rád, že fanúšikovia si všímajú mňa* ***namiesto toho, aby*** *povzbudzovali svoje mužstvo.*
(13c) *Bardejovčania v minulej sezóne vyhrali v Nitre 4:3,* ***keď ešte*** *v 85. minúte prehrávali 2:3.*
(13d) *Česká koruna oslabila na hodnotu 38,530 CZK/1EUR,* ***keď ešte*** *počas ranného obchodovania atakovala úroveň 38,350/1EUR.*

However, disappointment and disillusionment in the analysed utterances prevail overpleasant surprise from the unexpected consequence. This emotional-evaluative attitude is indicated mainly by the concessive operators *čo ako – akokoľvek – ani keď – i keby/aj keby* which are used mostly in belletristic texts. Each of them has a very specific meaning and collocations, but their common communicative function may be expressed with this relation: *hoci* + a negative emotional attitude of the text producer.

The conjunctional expresion *čo ako* is preferred in those instances, when the communication function of the utterance is to emphasize the contrast between an enormous effort and zero result. This conclusion is supported by the most frequent right-hand collocations of the conjunctional expression *čo ako* (verbs of volition *usilovať sa, snažiť sa, chcieť*) and intensifiers of action (*Ale* ***čo ako húževnato*** *by ste okolo neho magnetom čarovali, čo aj tým najsilnejším, aký len existuje, o akom sa Mesmerovi ani nesnívalo, sval sa nepohne.*).

The conjunction *akokoľvek* is functionally close to the conjunctional expression *čo ako* because it highlights the contrast between quite a high but not sufficient degree of a feature (***Akokoľvek budú principiálni****, ich jedinou možnosťou je momentálne politika kompromisu.*).

Utterances with the conjunctional expression *ani keď* present disappointment raising from the fact that something (some action or state)does not happen or occur even under extreme or extraordinary circumstances (*Keď ako štyridsaťpäťročný máte vzrast priemerného dieťa, nič na tom nezmeníte,* ***ani keď*** *budete mať deväťdesiatpäť.*)

Even a higher degree of contrast and disappointed expectations is reached in utterances with the operators *i keby/aj keby* presenting the fact that action/state does not happen/occur even if an unreal circumstance becomes a real one (*Nemôže sa naučiť algebru,* ***i keby*** *mal úžasného učiteľa. – Už nemôže nič povedať, nič dovysvetľovať,* ***aj keby*** *sa na to odvážila.)*.

The conjunctional expressions *navzdory tomu, že – bez ohľadu na to, že – nehľadiac na to, že* create utterances whose communication intention is to ensure the addressee aboutthe truth value of the proposition even if this seems to be surprising, improbable, incompatible with his/her experience or inconsistent with conventions. Their communication function can be described as *hoci* + insistent ensuring of the addressee about the truth value of the proposition.

(14a) *Ak poviete v New Yorku meno Andy,* ***nehľadiac na to, že*** *tam môže žiť 300 tisíc mužov s týmto menom, každý dodá Warhol.*
(14b) *Pri vstupe na štadión,* ***bez ohľadu na to, že*** *v minulosti bol najväčšou hviezdou domáceho trávnika, zaplatil si vstupné.*
(14c) *Felipe sa napokon definitívne rozhodol pre Letiziu,* ***navzdory tomu, že*** *nemala modrú krv.*

4.3.3 Finally, the paper focuses on concessive operators in utterances which have the same formal structure as "real" concessive utterances, but regarding the relation between their propositions, they do not match a negative implication *x, y* (thus it is not a relation between ineffective cause and unrealized consequence). These utterances are pragmatic and express speaker´s evaluating comment "I am aware of the fact that what I am going to say will raise doubts". In this way the speaker indicates that the proposition which he/she is going to say is not generally acceptable and he/she counts with a potential accusation of breaking utterance quality rules or principles of courtesy (***Hoci*** *je to paradoxné,...* ***– Akokoľvek*** *paradoxne to znie,...* ***– Hoci*** *to môže byť trochu pritiahnuté za vlasy,...* ***– Hoci*** *je to zvláštne,...* ***– Hoci*** *ide o otrepanú frázu,...* ***– Hoci*** *na prvý pohľad tieto dve veci spolu nesúvisia,...* ***– Akokoľvek*** *sa to zdá absurdné,...*). These utterances are close to the concessive relation because speaker´s doubtful evaluating comment turns out to be baseless in the further part of text. In my opinion, the category of quality motivates also utterances in which the concessive operator introduces a comment related to the speaker´s professional competence (***Hoc*** *ekológia nie je mojou parketou, napadá mi...,* ***– Hoci*** *nie som odborník na trestné právo, chcem uviesť...*).

5 Conclusion

The aim of this communicative-functional analysis is to depict dynamics within the micro-system of concessive operators. This dynamics is conditioned by different communication intentions of language users. Contemporary written Slovak has an extraordinarily miscellaneous and formally, stylistically, distributionally and pragmatically differentiated inventory of concessive operators. Dynamics in this microsystem is demonstrated by strengthening the position of a prototype representative and identifying the centre with synchronically most frequent operators on the one hand, and decline (even disappearance) of some synonymous operators on the other hand. Anothermanifestation of dynamics is a stylistic and pragmatic specification of operators which is not connected to the operator as a whole, but to its position in the utterance (initial or non-initial). Concessive operators are used mainly in belletristic and publicistic styles, less in a scientific style. The relation of negative implication which is the basis of concessive relation semantics is used pragmatically in many ways: in utterances with various communication functions such as statement, opinion, justification, evaluation, critique, defence against critique, ensuring of the addressee, surprise, astonishment, disappointment and disillusionment.

Author of the paper would like to thank to grant commission VEGA for supporting the project No. 1/0169/09 which allows realization of analyses presented in this paper.

References

[1] Dolník, J. – Bajzíková, E. (1998). Textová lingvistika. Bratislava: Stimul. 134 p.
[2] Dolník, J. (1999). *Základy lingvistiky*. Bratislava: Stimul. 228 p.
[3] Grepl. M. – Karlík, P. (1986). *Skladba spisovné češtiny*. Praha: Státní pedagogické nakladatelství. 474 p.
[4] Hirschová, M. (2006). *Pragmatika v češtine*. Olomouc: Univerzita Palackého. 243 p.
[5] Karlík, P. (1995). *Studie o českém souvětí*. Brno: Masarykova univerzita. 135 p.
[6] *Kolokace. Studie z korpusové lingvistiky*. (2006). Eds.: F. Čermák a M. Šulc. Praha: Nakladatelství Lidové noviny. 452 p.
[7] *Korpusová lingvistika. Stav a modelové přístupy*. (2006). Eds.: F. Čermák a R. Blatná. Praha: Nakladatelství Lidové noviny. 358 p.
[8] Müllerová, O. (1979). *Komunikativní složky výstavby dialogického textu*. Praha: Univerzita Karlova.. 161 p.
[9] *Morfológia slovenského jazyka*. (1966). Red. J. Ružička. 1. vyd. Bratislava: Vydavateľstvo Slovenskej akadémie vied. 895 p.
[10] Oravec, J. – Bajzíková, E. (1982). *Súčasný slovenský spisovný jazyk. Syntax*. Bratislava: Slovenské pedagogické nakladateľstvo. p. 171 – 175.

[11] Pauliny, E. (1981). *Slovenská gramatika*. Bratislava: Slovenské pedagogické nakladateľstvo. 288 p.
[12] Slovenský národný korpus – prim-4.0-public-all. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2009. Dostupný z WWW: http://korpus.juls.savba.sk.
[13] Slovenský národný korpus – prim-4.0-vyv. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2009. Dostupný z WWW: http://korpus.juls.savba.sk.
[14] *Slovník slovenského jazyka*. III. (1963). Vedecký redaktor Š. Peciar. Bratislava: Vydavateľstvo Slovenskej akadémie vied. 911 p.
[15] Slančová, D. (1994). *Praktická štylistika*. Prešov: Slovacontact. 178 p.
[16] Svozilová, H. (1986). Spojky a další spojovací prostředky. In: *Mluvnice češtiny. 2. Tvarosloví*. Praha: Academia. p. 214 – 227.
[17] Šulc, M. (2006). Frekvence javu v korpusu a dva typy jejího referenčního rámce. In: *Korpusová lingvistika: Stav a modelové přístupy*. Eds.: F. Čermák, R. Blatná. Praha: Nakladatelství Lidové noviny. p. 330 – 346.

Changes in Valency Structure of Verbs: Grammar vs. Lexicon*

Václava Kettnerová and Markéta Lopatková

Institute of Formal and Applied Linguistics
Charles University in Prague, Czech Republic

Abstract. In this paper, we deal with changes in valency structure of Czech verbs from a lexicographic point of view. We focus only on syntactic constructions that are related in principle to the same (generalized) situation. Changes in valency structure are understood as different mappings between individual participants of a generalized situation and valency slots, including their morphemic realization. We distinguish two types of changes in valency structure, so-called grammatical diatheses and semantic diatheses. We introduce a basic typology of potential changes in valency structure and we propose a method of the representation of these changes in the valency lexicon of Czech verbs VALLEX.

1 Motivation

Syntactic behavior of verbs is determined to a great extent by their lexical semantic properties. Prototypically, a single valency structure corresponds to a single meaning of verb. However, in many cases semantically related uses of verbs can be syntactically structured in different ways. E.g., the pairs of sentences in (1a)-(1b), (1a)-(2a) and (1b)-(2b) differ in their syntactic structure despite their obvious semantic similarity:

(1) a. *Peter loaded the truck with hay.* — b. *Peter loaded hay on the truck.*

(2) a. *The truck was loaded with hay.* — b. *Hay was loaded on the truck.*

Such uses of the verb *load* cannot be described by a single valency frame; however, separating four valency frames appears to be redundant with respect to the regularity in morphemic realizations of valency slots. Let us focus on the pairs of sentences (1a)-(2a) and (1b)-(2b). In these cases, (i) the information on the possibility of such change in valency structure of the verb *load* and (ii) the rule describing such change are sufficient for lexicographic description. Other changes in valency structure of verbs can be treated in a similar way under the condition that these changes are so regular that they can be captured by means of rules.

In this contribution, we deal with changes in valency structure of Czech verbs from a lexicographic point of view. We introduce and exemplify a basic typology of potential changes in valency structure of Czech verbs as they have appeared during the lexicographic processing language data (based on corpus evidence). Finally, we propose a method of representing these changes in a valency lexicon of Czech verbs.

* The research is carried under the MŠMT ČR project No. MSM0021620838 and partially under the MŠMT grant No. LC536 and GA UK grant No. 7982/2007.

Basic approaches to changes in valency structure. In Czech linguistics, the study of syntactic constructions characterized by changes in valency structure of verbs from the syntactic point of view started in the late sixties, mainly under the influence of Russian linguistics, esp. [1, 3, 6]. The terms hierachization, diathesis or conversion were introduced in Czech and Slovak grammars, see esp. [7, 8, 15, 21] and [11]. Roughly speaking, such terms refer to change in mutual assignment of semantic participants and (surface) syntactic positions, while the real situation expressed by sentences remains the same.

In American linguistics, there are three basic approaches to changes in valency structure of verbs, (i) structurally based approaches represented mainly by transformational-generative grammars, esp. [4, 5], (ii) lexically based approaches focusing on the relation between lexical semantic properties of verbs and their syntactic behavior, esp. [12], and (iii) constructionally based approaches based on the assumption that difference in syntactic forms marks the difference in meaning, esp. [2, 10].

Here we focus on the description of changes in valency structure of verbs in the theoretical framework of the Functional Generative Description (FGD), see esp. [20]. The valency theory of FGD, esp. [16], was applied to a large number of data in building the Prague Dependency Treebank, PDT 2.0[1] and the valency lexicon of Czech verbs, VALLEX[2] [13]. We attempt to propose an adequate framework for description of changes in valency structure of verbs which can be applied in lexicographic processing of language data.

2 Basic typology of changes in valency structure of verbs

In our typology of changes in valency structure of verbs, the concept of situation plays a key role. The **(generalized) situation** represents a class of abstract situations characterized by a particular set of semantic participants.[3] In the present paper, we focus only on those syntactic constructions that relate to the same (generalized) situation. Such a situation is expressed by a single verb lexeme and it is characterized by an identical set of semantic participants. Changes in valency structure are understood as different mappings between individual semantic participants of a generalized situation and their surface syntactic positions, including their morphemic realization. We distinguish two types of changes in valency structure, so-called grammatical diatheses (g-diatheses) and semantic diatheses (s-diatheses).

2.1 Grammatical diatheses

G-diatheses represent pairs of related syntactic constructions that prototypically satisfy the following criteria:

[1] http://ufal.mff.cuni.cz/pdt2.0/

[2] http://ufal.mff.cuni.cz/vallex/2.5/

[3] See also type situation [8, 22] or semantic event. Semantic participants roughly correspond to semantic roles here.

I. Verbs in the marked construction are prototypically morphologically marked with respect to the grammatical category of voice. Their forms typically either consist of auxiliaries and non-finite form of lexical verbs or they have reflexive forms.
II. The mapping between semantic participants of a generalized situation and valency slots remains unchanged, their number and type are identical as well. Changes in valency frames are typically connected with a choice of a particular valency member for the subject syntactic positions; these changes are limited to morphemic realizations of individual valency slots.

G-diatheses primarily represent a language means that enables the speaker to choose a particular semantic participant of a generalized situation for the syntactically prominent position of (surface) subject. In the marked case, the valency member ACT (Actor, corresponding to the semantic participants of generalized situation such as Agent, Initiator, Causator, Bearer of Action, etc.) is prototypically shifted from the subject syntactic position into a less prominent surface position; eventually, it cannot be expressed on the surface syntactic level at all (as in deagentive g-diathesis, see e.g. [9]). Another semantic participant of a generalized situation (typically having the form of accusative) is shifted into the subject syntactic position, as in (1a)-(2a) repeated below.[4] Under certain conditions, a 'subject-less' construction occurs (see example (7b) below).

(1) a. *Peter*.ACT *loaded the truck*.PAT *with hay*.EFF

(2) a. *The truck*.PAT *was loaded with hay*.EFF (*by Peter*.ACT)

G-diatheses can be illustrated by the scheme in Figure 1, the asymmetry concerns the different mappings between a set of valency members and their surface positions.

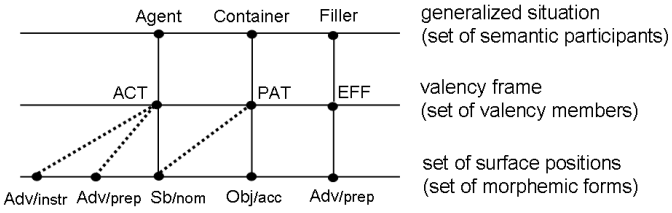


Fig. 1. Mapping between semantic participants of a generalized situation and their surface syntactic positions for passive diathesis as a typical g-diathesis (for the verb *naložit* 'to load').

We assume that changes in the valency structure of verbs characteristic of g-diatheses arise from the special verbal meanings. These verbal meanings are reflected as values of relevant verbal grammatemes in FGD (grammatemes represent tectogrammatical correlates of the morphological categories, see [14, 19]).

[4] We mark the valency members with labels (so-called functors) ACT, PAT, EFF etc. in accordance with FGD (and with VALLEX in particular).

2.2 Semantic diatheses

S-diatheses are characterized by changes in number and type of valency slots, while the (generalized) situation still remains unchanged. Furthermore, verbs are not morphologically marked with regard to voice. Contrary to g-diatheses, it is not apparent which of the related constructions should be understood as unmarked ones and which as marked ones, see also [8].

Moreover, s-diatheses are typically associated with coherent semantic classes of verbs, as in sentences (1a)-(1b) (see also, e.g., *spray/load* verbs in [12]).

(1) a. *Peter*.ACT-Agent *loaded the truck*.PAT-Container
with hay.EFF-Filler
b. *Peter*.ACT-Agent *loaded hay*.PAT-Filler
on the truck.DIR-Container

In Czech grammars, s-diatheses are described as hierarchizations without marked voice [8], as objective diatheses [11], or some of them are treated as examples of the so-called decauzativization [11].

S-diatheses can be illustrated by the scheme in Figure 2, the asymmetry concerns the different mappings between a set of semantic participants of a generalized situation and a set of valency members.

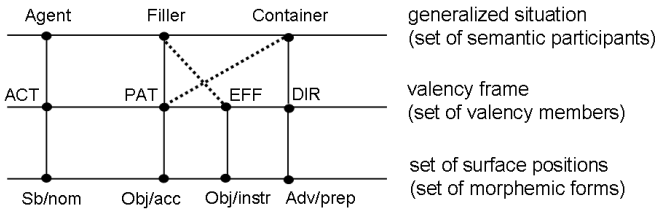


Fig. 2. Mapping between semantic participants of a generalized situation and their surface syntactic positions for Container-Filler diathesis (for the verb *naložit* 'to load').

As to the possibility of combining g- and s-diatheses, diatheses of different types are mutually combinable; i.e., having a marked construction with respect to a g-diathesis, a particular s-diathesis rule may be subsequently used (if applicable for the given verb), and conversely, see ex. (1)-(2) in Section 1. However, mutually combining diatheses of the same type is very restricted.[5]

Distinguishing between g-diatheses and s-diatheses is motivated by the needs of lexicographic work. We will see later that in case of **g-diatheses**, the changes in valency

[5] E.g., *Když se dostane přidělena pracovna, to se to pracuje.* — Eng. If a new study is allocated, it is easy to work (example from [9]).

frames are regular enough to be treated within a single verbal lexical unit – general rules in the grammar component and information on their applicability to individual lexical units in the data component of the lexicon are sufficient. However, for **s-diatheses**, we propose to set separate lexical units interlinked with general rules identifying a relevant type of s-diathesis. This solution results from the corpus evidence that changes in valency structure of verbs are diverse even within an individual type of s-diatheses.

3 Representation of G-diatheses

In this section, we introduce a way of capturing g-diatheses in the valency lexicon VALLEX. In our approach, g-diatheses are described by means of general fine-grained rules in the grammar component of the valency lexicon. All applicable g-diatheses are listed for each verbal lexical unit separately in a special attribute in the data component of the lexicon.

Our method will be demonstrated on the passive diathesis as a prototypical g-diathesis. **Deagentive diathesis**, **recipient diathesis**, **resultative diathesis** and **mediopassive diathesis**, see esp. [19], can be described in the same way. In addition, we consider also **reciprocity** as a phenomenon that can be treated in a similar way (within FGD, reciprocity and the possibility of its representation have been broadly studied by Panevová, esp. [17]).[6]

3.1 Passive diathesis

Passive diathesis is a relation between two syntactic constructions in which the marked one contains the auxiliary verb *být* ‘to be’ and the past participle of a lexical verb. We propose the following representation of passive diathesis in the valency lexicon:

(i) In the **data component**, a single lexical unit is represented by an (unmarked) valency frame. If a given lexical unit can be subject to passive diathesis, then its applicability is indicated in the special attribute ‘diathesis-pass’.
(i) In the **grammar component**, a general rule describing regular changes in a valency frame for this diathesis is stored.

For example, a lexical unit for the transitive verb *postavit* ‘to build’ has three valency slots in its valency frame: obligatory ACT (Actor, in nominative in the unmarked construction), obligatory PAT (Patient, in accusative) and optional ORIG (Origin, expressed as the prepositional group *z* ‘from/of’ plus genitive). In the marked construction, ACT is realized either as instrumental or as prepositional group *od* ‘by’ plus genitive, and PAT is expressed as nominative (morphemic realization of ORIG remains unchanged):

(3) a. *David*.ACT$_{nom}$ *postavil kůlnu*.PAT$_{acc}$ *ze dřeva*.ORIG$_{z+gen}$
Eng. David.ACT built a shed.PAT from wood.ORIG
b. *Kůlna*.PAT$_{nom}$ *byla postavena ze dřeva*.ORIG$_{z+gen}$ (*Davidem / od Davida*.ACT$_{instr,od+gen}$)
Eng. A shed.PAT was built from wood.ORIG (by David.ACT)

[6] **Causative constructions** are another candidates that can be taken into account for this type of representation.

Passive diathesis for verbs with valency member expressed by accusative. Passive diathesis concerns verbs with at least two semantic participants of a generalized situation and thus at least two valency slots, prototypically ACT in nominative and PAT in accusative. Valency frame for the marked member of the diathesis can be described by the following rule **Pass.r1.PAT**, see Table 1.

It should be stressed here that all information captured in valency frame remains unchanged, unless a change is explicitly mentioned by the rule Pass.r1.PAT; i.e., if a valency frame contains a member or morphemic form that is not cited in the rule, then it is preserved also in a derived valency frame.

Pass.r1.PAT	Unmarked	Marked	Note
verbal grammateme	diathesis-pass: 0	diathesis-pass: pass	(1)
valency frame	ACT_{nom}	$\text{ACT}_{instr,od+gen}$	(2)
	PAT_{acc}	PAT_{nom}	(3)
	$\text{PAT}_{var,inf,dcc}$	$\text{PAT}_{excluded}$	(4)
	? $\text{EFF}_{jako+acc}$	? $\text{EFF}_{jako+nom}$	(5)

Table 1. Pass.r1.PAT rule for the passive diathesis.

Commentary on the Pass.r1.PAT rule:

(1) The passive diathesis is represented by the verbal grammateme 'diathesis-pass'; its value for the unmarked member of the pair is '0', for the marked member it is 'pass'.

(2) In the marked construction, ACT is shifted from the prominent subject syntactic position into the adverbial position. This change is accompanied by the change of morphemic realization of ACT from nominative into instrumental or into the prepositional case *od* 'by'+genitive.

(3) The valency member PAT (expressed by accusative) is selected for the prominent surface syntactic position of subject for the marked member of the passive diathesis. Its morphemic form is changed into nominative.

(4) If the PAT valency member may be expressed also by other morphemic forms such as infinitive (abbr. *inf*), dependent content clause (*dcc*) or another preposition or prepositionless case (*var*) (mentioned below as 'unaccusative variants'), all these possible morphemic variants are excluded in the marked frame. PAT expressed by unaccusative forms is treated with Pass.r2.PAT rule, see below.

(5) If there is a slot for EFF in the unmarked frame with the form *jako* 'as'+accusative, then its form is changed into *jako* 'as'+nominative.

Note on agreement: Verbal categories of person, number and gender agree with ACT in nominative in the unmarked construction, whereas a verb in the marked construction has agreement with PAT in nominative.

For example, by applying Pass.r1.PAT rule to the unmarked valency frame for the verb *postavit* 'to build', see ex. (3a)-(3b), we obtain the following valency frame describing the marked syntactic construction:

$$\text{ACT}_{nom}\ \text{PAT}_{acc}\ \text{ORIG}_{z+gen} \Rightarrow_{\text{Pass.r1.PAT}} \text{ACT}_{instr,od+gen}\ \text{PAT}_{nom}\ \text{ORIG}_{z+gen}$$

The change in the realization of EFF expressed with *jako* 'as'+accusative may be exemplified by the verb *hodnotit* 'to assess'. See the unmarked and marked valency frames and their realizations in sentences (4a)-(4b) (note also the reduction of possible morphemic forms for PAT in (4b)):

$$\text{ACT}_{nom}\ \text{PAT}_{acc,var,inf,dcc}\ \text{EFF}_{jako+acc,na+acc} \Rightarrow_{\text{Pass.r1.PAT}} \text{ACT}_{instr,od+gen}\ \text{PAT}_{nom}\ \text{EFF}_{jako+nom,na+acc}$$

(4) a. *Učitelé*.ACT$_{nom}$ *hodnotili jeho práci*.PAT$_{acc}$ *jako nedostatečnou*.EFF$_{as+acc}$
Eng. The teachers.ACT assessed his paper.PAT as poor.EFF
b. *Jeho práce*.PAT$_{nom}$ *byla hodnocena učiteli*.ACT$_{instr}$ *jako nedostatečná*.EFF$_{as+nom}$
Eng. His paper.PAT was assessed as poor.EFF by his teachers.ACT

For some verbs with at least three valency members, the accusative position may be labeled with other functors, namely ADDR (for Addressee) or EFF (for Effect),[7] see (5a)-(5b) and (6a)-(6b). The changes in valency structure of these verbs are captured by analogous rules Pass.r1.ADDR and Pass.r1.EFF.

(5) a. *Sekretářka*.ACT$_{nom}$ *ředitele*.ADDR$_{acc}$ *upozornila*, (*že má podepsat smlouvu*).PAT$_{dcc}$
Eng. The secretary.ACT has reminded the director.ADDR (to sign the contract).PAT
b. *Ředitel*.ADDR$_{nom}$ *byl upozorněn sekretářkou*.ACT$_{instr}$, (*že má podepsat smlouvu*).PAT$_{dcc}$
Eng. The director.ADDR has been reminded by his secretary.ACT (to sign the contract).PAT

(6) a. *Zadržený*.ACT$_{nom}$ *řekl vyšetřovateli*.ADDR$_{dat}$ *lež*.EFF$_{acc}$
Eng. The detained man.ACT said to the interrogator.ADDR a lie.EFF
b. *Vyšetřovateli*.ADDR$_{dat}$ *byla* (*zadrženým*.ACT$_{instr}$) *řečena lež*.EFF$_{nom}$
Eng. A lie.EFF was said to the interrogator.ADDR (by the detained man.ACT)

Passive diathesis for verbs with valency member expressed by 'unaccusative' forms. Furthermore, passive diathesis can be applied to verbs with valency members realized by 'unaccusative' forms, see ex. (7a)-(7b):

(7) a. *Radní*.ACT$_{nom}$ *o té záležitosti*.PAT$_{o+loc}$ *rozhodli včera*.
Eng. The councilors.ACT decided the matter.PAT yesterday.
b. *O té záležitosti*.PAT$_{o+loc}$ *bylo* (*radními*.ACT$_{instr}$) *rozhodnuto včera*.
Eng. The matter.PAT was decided (by councilors.ACT) yesterday.

Changes in valency frame are described by the following rule **Pass.r2.PAT**, see Table 2. Again, except for the changes explicitly mentioned in the rule, all other information captured in a valency frame remains unchanged.

[7] We leave aside the functors DPHR (for Dependent Part of Phraseme) and CPHR (Part of Compound Predicate) here.

Pass.r2.PAT	Unmarked	Marked	Note
verbal grammateme	diathesis-pass: 0	diathesis-pass: pass	(1)
valency frame	ACT_{nom}	$\mathrm{ACT}_{instr,od+gen}$	(2)
	$\mathrm{PAT}_{var,inf,dcc}$	$\mathrm{PAT}_{var,inf,dcc}$	(3)
	? PAT|ADDR|EFF_{acc}	? PAT|ADDR|$\mathrm{EFF}_{excluded}$	(4)

Table 2. Pass.r2.PAT rule for the passive diathesis.

Commentary on the Pass.r2.PAT rule:
(1) and (2) See the Commentary on the Pass.r1 rule.
(3) The 'unaccusative' morphemic realization of PAT[8] remains unchanged. If PAT is realized by infinitive or dependent content clause, it is shifted into the subject syntactic position. Applying the given rule to PAT expressed by prepositional case or prepositionless case (with the exception of accusative), 'subject-less' sentence is created.
(4) The possible accusative realization of any valency slot is excluded. If no other morphemic variant remains, the given valency member cannot be realized in a surface sentence,[9] see also ex. (8c).
Note on agreement: In the marked construction, verbs have incongruent agreement with 3^{rd} sg. neutr.

Let us exemplify the application of Pass.r2.PAT rule to the valency frame of the verb *rozhodnout* 'to decide', see also sentences (7a)-(7b):

$$\mathrm{ACT}_{nom}\ \mathrm{PAT}_{o+loc,dcc} \Rightarrow_{\text{Pass.r2.PAT}} \mathrm{ACT}_{instr}\ \mathrm{PAT}_{o+loc,dcc}$$

Verbs allowing for two passive constructions. There are verbs allowing for two passive constructions. First, such verb has a valency member that may be realized both as accusative and 'unaccusative' form (e.g., the verb *hodnotit* 'to asses', see ex. (4)) – then both types of rules are applicable to this valency member (Pass.r1.PAT or Pass.r2.PAT for the verb *hodnotit* 'to asses'). The second case is represented by verbs with at least three semantic participants of generalized situations. Such verbs have at least three valency members (prototypically realized as nominative, accusative and 'unaccusative').[10] Again, both types of rules may be used – they are applied to two different valency members depending on the choice of subject. We exemplify this by the verb *žádat* 'to ask', see sentence (8a) for the unmarked case, (8b) for the Pass.r1.ADDR rule and (8c) for the Pass.r2.PAT rule:

$$\mathrm{ACT}_{nom}\ \mathrm{ADDR}_{acc}\ \mathrm{PAT}_{o+acc,inf,dcc} \Rightarrow_{\text{Pass.r1.ADDR}} \mathrm{ACT}_{instr,od+gen}\ \mathrm{ADDR}_{nom}\ \mathrm{PAT}_{o+acc,inf,dcc}$$

$$\mathrm{ACT}_{nom}\ \mathrm{ADDR}_{acc}\ \mathrm{PAT}_{o+acc,inf,dcc} \Rightarrow_{\text{Pass.r2.PAT}} \mathrm{ACT}_{instr,od+gen}\ \mathrm{ADDR}_{general}\ \mathrm{PAT}_{o+acc,inf,dcc}$$

[8] The analogous rules are set for ADDR and EFF.
[9] This case results in so called generalized valency member in FGD, see [18].
[10] The verb *učit* 'to teach' with two valency members expressed in accusative represents a rare exception.

As the accusative is the only possible realization of ADDR in the unmarked valency slot (and accusative is excluded in the marked valency frame according to Pass.r2.PAT rule), the ADDR valency slot cannot be realized in the surface sentence, see ex. (8c).

(8) a. *Novináři*.ACT$_{nom}$ *vládu*.ADDR$_{acc}$ *žádali,* (*aby byly zveřejněny výsledky*).PAT$_{dcc}$
Eng. The journalists.ACT asked the government.ADDR (to publish the results).PAT
b. *Vláda*.ADDR$_{nom}$ *byla* (*novináři*.ACT$_{instr}$) *žádána,* (*aby byly zveřejněny výsledky*).PAT$_{dcc}$
Eng. The government.ADDR was asked (by the journalists.ACT) (to publish the results).PAT
c. *Novináři*.ACT$_{instr}$ *bylo opakovaně žádáno,* (*aby byly zveřejněny výsledky*).PAT$_{dcc}$ (general ADDR)
'(by) journalists - was - repeatedly - asked - to - publish - results' Eng. The publication of the results was repeatedly asked (by the journalists).

4 Representation of S-diatheses

In this section, we focus on s-diatheses and their adequate representation in the valency lexicon VALLEX. To recapitulate, s-diathesis is a relation between two (or more) syntactic constructions describing a same generalized situation. These constructions refer to the same (polysemous) verb lexeme, however, the mappings between individual semantic participants of the generalized situation and valency slots is different. As a consequence, not only morphemic realization but also number, type and obligatoriness of valency members may differ. In contrast to g-diatheses, morphological categories of the given verb typically remain unchanged.

Let us demonstrate our approach on the Container-Filler diathesis as a prototypical s-diathesis. Other s-diatheses can be captured in the same way (selected examples are listed below).

4.1 Container-Filler diathesis

Container-Filler diathesis[11] can be exemplified by sentences (9a)-(9b) (note that 'negative' variant can be also distinguished).

(9) a. *Petr*.ACT$_{nom}$-Agent *naložil vůz*.PAT$_{acc}$-Container *senem*.EFF$_{instr}$-Filler
Eng. Petr.ACT-Agent loaded the truck.PAT-Container with hay.EFF-Filler
b. *Petr*.ACT$_{nom}$-Agent *naložil seno*.PAT$_{acc}$-Filler

[11] This type of diathesis counts among a group of 'co-occurrence diathesis' in [8]; see also 'spray/load alternation' in [12]. We adopt a labeling based on semantic participants involved in the diatheses as we consider it more transparent.

na vůz.DIR-Container
Eng. Petr.ACT-Agent loaded hay.PAT-Filler
on the truck.DIR-Container

These two sentences describe in principle the same generalized situation with three semantic participants – Agent (who causes the action described by the given verb), Filler (substance or entity whose location is changed) and Container (location where Filler is moved). Despite the single set of semantic participants of the generalized situation, this situation can be structured in a different way. While Agent is realized as ACT in both cases, there are two possibilities for Filler and Container: (i) either Container is mapped onto PAT (in accusative) and Filler is mapped onto EFF valency slot (in instrumental), as in (9a); (ii) or Filler occupies the PAT slot (in accusative) and Container is structured as Directional modification DIR, as in (9b) (see also Figure 2 in Section 2.2).

The most studied semantic property of this diathesis deals with a partitive / holistic effect. The semantic participant of the generalized situation realized as PAT in accusative typically receives holistic interpretation; i.e., in Container-Filler diathesis either Container (9a) or Filler (9b) is understood as completely affected by the action expressed by the verb *naložit* 'to load'.

Contrary to g-diatheses, the changes in valency frames accompanying s-diatheses are not regular enough: individual verbs exhibit many irregularities in their valency characteristics even within a single type of s-diathesis (see below for the examples).

For the purpose of the valency lexicon VALLEX, we propose the following representation of s-diatheses:

(i) In the **data component**, we establish a set of two lexical units within one lexeme – each member of s-diathesis is represented by a separate lexical unit with its own valency frame. These lexical units are interlinked via the type of s-diathesis (captured in a special attribute 's-diathesis').
(ii) In the **grammar component**, a general rule describing possible mappings between semantic participants of a generalized situation and individual valency slots is provided, see Table 3.

Container-Filler	**Agent**	**Filler**	**Container**	examples
Filler ∼ **PAT**	ACT	PAT	DIR	*naložit seno na vůz* *doplnit cukr do cukřenky* *nasypat mouku do pytle* *(na)točit vodu (do kýble)*
Container ∼ **PAT**	ACT	EFF	PAT	*naložit vůz senem* *doplnit cukřenku cukrem/o cukr*
	ACT	—	PAT	*nasypat pytel *moukou* *(na)točit kýbl *vodou*

Table 3. General rule for the Container-Filler diathesis (see the translations below).

The dissimilarities in the Container-Filler diathesis concern number, type, and morphemic realization of complements as well:

- Whereas the set of semantic participants of the generalized situation is the same (Agent, Filler, Container) and prototypically all of them can be realized as valency members, this does not hold for some verbs (e.g., *nasypat mouku do pytle* 'to put flour into the sack' but *nasypat pytel *moukou* 'to put the sack *with flour').
- Whereas directional valency member that realizes Container participant is prototypically obligatory (e.g., *doplnit cukr do cukřenky* 'to add sugar to the sugar bowl'), there are verbs with only typical directional valency member (e.g., *točit vodu (do kýble)* 'to draw water (to the bucket)').
- Morphemic realizations of a particular valency member may differ with individual verbs (e.g., *doplnit cukřenku cukrem / o cukr* 'to replenish the sugar bowl with sugar').

4.2 Examples of other S-diatheses

While g-diatheses are intensively studied in Czech linguistics, there is only a limited number of studies of phenomena referred here to as s-diatheses, see esp. [8]. Let us exemplify here at least several frequent s-diatheses in Czech which can be captured in the valency lexicon in a similar way as the Container-Filler diathesis:

Surface-Cover diathesis (positive or negative)
Jana si očistila bláto.PAT-Cover *z bot*.DIR-Surface
Eng. Jane cleaned the mud.PAT-Cover off her shoes.DIR-Surface
— *Jana si očistila boty*.PAT-Surface *od bláta*.ORIG-Cover
Eng. Jane cleaned her shoes.PAT-Surface of the mud.ORIG-Cover

Material-Product diathesis (positive or negative)
Kadeřník jí učesal vlasy.PAT-Material *do drdolu*.EFF-Product
Eng. Hairdresser arranged her hair.PAT-Material into a bun.EFF-Product
— *Kadeřník jí učesal z vlasů*.ORIG-Material *drdol*.PAT-Product
Eng. Hairdresser arranged a bun.PAT-Product from her hair.ORIG-Material

Source-Substance diathesis
Slunce.ACT-Source *vyzařuje teplo*.PAT-Substance
Eng. The sun.ACT-Source radiates heat.PAT-Substance
— *Teplo*.ACT-Substance *vyzařuje ze slunce*.DIR-Source
Eng. Heat.ACT-Substance radiates from the sun.DIR-Source

Object-Direction diathesis ('from where', 'through' or 'to where')
Marta vylezla kopec.PAT-Object
Eng. Martha climbed the mountain.PAT-Object
— *Marta vylezla na kopec*.DIR-Direction
Eng. Martha climbed up the mountain.DIR-Direction

Direction-Location diathesis
Matka umístila dítě do jeslí.DIR-Direction
Eng. Mother put her child into a nursery school.DIR-Direction
— *Matka umístila dítě v jeslích*.LOC-Location
Eng. Mother put her child into a nursery school.LOC-Location

Agent-Location diathesis

Včely.ACT-Agent *se rojí na zahradě*.LOC-Location
Eng. Bees.ACT-Agent are swarming in the garden.LOC-Location
— *Zahrada*.ACT-Location *se rojí včelami*.MEANS-Agent
Eng. The garden.ACT-Location is swarming with bees.MEANS-Agent

Conclusion

For lexicographic description of verbal valency, it is necessary to specify (i) valency frame of each lexical unit, (ii) information on the applicability of a particular set of rules describing the possible diatheses, and (iii) precise formulations of rules. Information (i) and (ii) are stored in the data component whereas (iii) is stored in the grammar component of the valency lexicon.

We distinguish two types of changes in valency structure, which are referred to as g-diatheses and s-diatheses. G-diatheses are prototypically characterized by morphologically marked form of verb in the marked construction, while the mapping between semantic participants of a generalized situation and valency slots remains unchanged, their number and type are identical (the changes in valency frames are limited to morphemic realizations of individual valency slots). On the other hand, s-diatheses are characterized by changes in number and types of valency slots. They are typically limited to verbs of certain semantic classes.

Distinguishing between g-diatheses and s-diatheses in the valency lexicon VALLEX is motivated by the needs of lexicographic work. In case of g-diatheses, the changes in valency frames are regular enough to be treated in the form of general rules (in the grammar component) and as a single verbal lexical unit (for both syntactic constructions) marked with the possibility of a particular type of diathesis. For s-diatheses, separate lexical units are established and interlinked with general rules identifying a relevant type of s-diathesis. This solution reflects the corpus evidence that changes in valency structure of verbs are diverse even within an individual type of s-diathesis.

References

[1] Apresjan, J. D. (1974). *Leksicheskaja semantika. Sinonimicheskie sredstva jazyka*. Nauka, Moskva.
[2] Borer, H. (2005). *The Normal Course of Events*. Oxford University Press, Oxford.
[3] Cholodovič, A. A. (1970). Zalog. Kategoria zaloga. In *Materialy konferencii*, pages 2–26, Leningrad.
[4] Chomsky, N. A. (1957). *Syntactic Structures*. Mouton, The Hague.
[5] Chomsky, N. A. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge.
[6] Chrakovskij, V. S., editor (1977). *Problemy lingvisticheskoj tipologii i struktury jazyka*. Nauka, Leningrad.
[7] Daneš, F. (1968). Some Thoughts on the Semantic Structure of the Sentence. *Lingua*, 21:55–69.
[8] Daneš, F. (1985). *Věta a text: studie ze syntaxe současné češtiny*. Academia, Praha.

[9] Daneš, F., Grepl, M., and Hlavsa, Z., editors (1987). *Mluvnice češtiny 3*. Academia, Praha.
[10] Goldberg, A. E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.
[11] Grepl, M. and Karlík, P. (1998). *Skladba češtiny*. Votobia, Olomouc.
[12] Levin, B. C. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago and London.
[13] Lopatková, M., Žabokrtský, Z., and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha.
[14] Mikulová, M. et al. (2006). Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report TR-2006-30, ÚFAL MFF UK, Prague.
[15] Ondrejovič, S. (1989). *Medzi slovesom a vetou*. Jazykovedné štúdie. Veda, Bratislava.
[16] Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Academia, Praha.
[17] Panevová, J. (2007). Znovu o reciprocitě. *Slovo a slovesnost*, 68:91–100.
[18] Panevová, J. and Řezníčková, V. (2001). K možnému pojetí všeobecnosti aktantu. In Hladká, Z. and Karlík, P., editors, *Čeština – univerzália a specifika 3*, pages 139–146. Masarykova Universita, Brno.
[19] Panevová, J. et. al (manuscript). *Syntax současné češtiny (na základě anotovaného korpusu)*. Nakladatelství Karolinum, Praha.
[20] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.
[21] Štícha, F. (1984). *Utváření a hierarchizace struktury větného znaku*. Univerzita Karlova, Praha.
[22] Uspenskij, V. A. (1977). K ponjatiju diatezy. In Chrakovskij, V. S., editor, *Problemy lingvisticheskoj tipologii i struktury jazyka*, pages 65–84. Leningrad.

Corpus-Based Analysis of Lexico-Grammatical Patterns (on the Corpus of Letters of N. V. Gogol)*

Maria Khokhlova[1,2,] and Victor Zakharov[1,2]

[1] Saint-Petersburg State University
[2] Institute of Linguistic Studies, Russian Academy of Sciences, Russia

Abstract. The paper is concerned with collocation extraction by means of statistical methods combined with certain syntactic models. Among collocations there is a special type of them which is usually called colligations. It could be defined as lexico-grammatical patterns, or collocations with regard to syntactical formulae. There's a system known as Sketch Engine representing itself a corpus tool which takes as input a corpus of any language and corresponding grammar patterns. The system gives information about a word's collocability on concrete dependency models, and generates lists of the most frequent phrases for a given word based on appropriate models. The present paper deals with collocation extraction on the corpus letters of N.V. Gogol by means of the mentioned system, discusses the results obtained. The results show that word sketches and information about collocation behaviour could facilitate lexicographic work with the Russian language.

1 Introduction

Nowadays in modern linguistics corpora have turned to be vital tools for linguistic studies and solution for applied tasks. The application of corpora methods to the analysis of lexical collocability enables to write grammars and compile dictionaries of a new type, dictionaries of collocations, idioms etc. With arrival of text corpora and corpus linguistics lexicographers and other linguists have gained an opportunity to look at big collections of word usage. Corpora do not only help to study lexical units in context but also to get data on word frequency, frequency of lexemes, grammatical categories, their collocability etc.

Although the above mentioned corpora opportunities are very useful, there is a need of another kind of software for further improvement of linguistic research as it is impossible to process huge amount of linguistic data manually. It can be described as an additional system between a corpus and its users (linguists) which can process significant language data.

The objective of such a system is to provide lexicographers with sufficient lexical material and tools for getting information about a word's collocability and to generate lists of the most frequent phrases for a given word, and then to classify them for appropriate syntactic models.

* The project is supported by Visegrad Scholarship "Syntactic and statistical models of phrases in Russian" (Contract Number #50810256).

2 Sketch engine

Such a system known as Sketch Engine was developed for a number of languages by British and Czech scholars (see Sketch Engine project[1]). The Sketch Engine combines approaches of both traditional linguistics (syntax) and statistics. It is widely used by scholars in grammar writing and dictionary compiling (Oxford University Press, Cambridge University Press, Collins, Macmillan etc.).

The Sketch Engine system is developed for a number of European languages (English, Irish, Spanish, Italian, German, Portuguese, Slovene, French, Czech) and for the Chinese and Japanese languages. This is a corpus tool which takes as input a corpus of any language and corresponding grammar patterns and which generates word sketches for words of that language. Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour [1]. One can understand word sketches as typical phrases determined on the one hand by syntax that restricts words' collocability in a given language and on the other hand by probability closely related to word usage.

3 Word sketch grammar (rules for the Russian language)

The paper presents results of writing rules for the Russian language for the system Sketch Engine and testing them on the corpus of letters of N. V. Gogol (a famous Russian writer of the XIXth century).

During our work we have written grammatical rules that take into account syntactic constructions of the Russian language based on the morphologically tagged corpus in terms of grammar of Sketch Engine. On the basis of these rules and statistical measures the programme generates tables with word sketches for a key word.

While writing rules we used regular expressions and query language IMS Corpus Workbench. The system searches for tags which correspond to word forms. For example, tag *Ncfpnn* means common noun (*Nc*) female gender (*f*) plural (*p*) noun case (*n*).

Below there is an example of grammatical rules for the phrases «adjective+noun»:

```
*DUAL
=a_modifier/modifies
2:"A....n." (([word=","]|[word="и"]|[word="или"]) [tag="A....n."]){0,3} 1:"N...n."
2:"A....g." (([word=","]|[word="и"]|[word="или"]) [tag="A....g."]){0,3} 1:"N...g."
2:"A....d." (([word=","]|[word="и"]|[word="или"]) [tag="A....d."]){0,3} 1:"N...d."
2:"A....a." (([word=","]|[word="и"]|[word="или"]) [tag="A....a."]){0,3} 1:"N...a."
2:"A....i." (([word=","]|[word="и"]|[word="или"]) [tag="A....i."]){0,3} 1:"N...i."
2:"A....l." (([word=","]|[word="и"]|[word="или"]) [tag="A....j."]){0,3} 1:"N...l."
```

[1] http://www.sketchengine.co.uk

Above mentioned rules take into account all such phrases, e.g. nouns and adjectives in the same case with conjunctions «и» (“and”), «или» (“or”), comma or adjectives between them within the distance of 3 words. The numeral 1 stands for the key word (for instance, *1:"N...n."*) and the numeral 2 indicates a collocate (for instance, *2:"A....n."*).

Originally these rules were written on the basis of existing rules for English and Czech [2].

Then we have written the second variant of word sketches rules within the approach of Vladimir Benko for the Slovak National Corpus [3].

Its distinctive feature is that these rules describe all phrases found in a corpus. For example, “verb + any word” (see below):

```
=Vb X/X Vb
2:[tag="V.*"] 1:[tag!="SENT"]
1:[tag!="SENT"] 2:[tag="V.*"]
```

The second line means that there will be found all phrases for any word (if it isn’t a punctuation mark that has its own tag in the corpus) with a verb. The rule in the third line describes the same phrases but a verb is to the right of the key word.

It should be remarked that this approach has its advantage as word sketches are generated for any word (because very often morphological ambiguity or mistakes of automatic tagging prevent from giving objective results when a given word is specified).

As statistical tools in the system there are used the following measures of association: log-likelihood, MI, t-score, MI3, widely used in linguistics for the computation of the strength of syntagmatic relations.

4 Collocation extraction on the corpus of letters of N. V. Gogol

We have used a collection of letters of N.V. Gogol as test material (about 500 thousand tokens) [4]. Texts were morphologically tagged by the programme TreeTagger [5]. During this work we have tested the rules and got information about collocability on the corpus data.

Below one can see examples of phrases with the key word «маминька» (“mum”).

Fig. 1. Example of phrases with the key word «маминька» (“mum”)

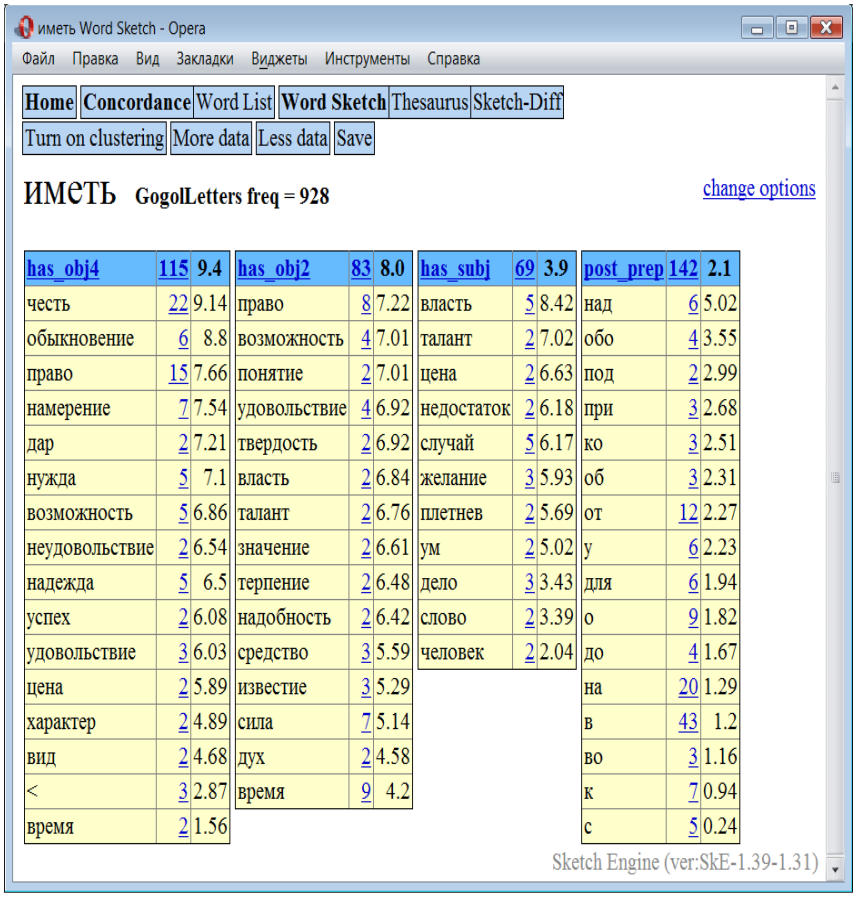
The first column represents left collocates of the key word «маминька» (“mum”).

The adjective «Дражайшая» (“dear”, archaic) (the 5th line of a modifier pattern) was selected as collocate due to the mistake of lemmatization (a form in female gender was ascribed to the word, moreover, as a proper name, this is supported by its uppercase spelling). The same mistake one can find with the word «Папинька» (“dad”, archaic).

The second column represents frequency of the given collocate. The third column shows the value of statistical measure according to which the collocate was selected, it is connected by hyperlink with concordance.

It should be stressed that parts of one phrase can be delimited by other words. For example, for the noun «маминька» there is a verb «обещаться» (“to promise”), that is located at the distance of two words to the right: *Также ежели б еще прислали чего-нибудь из съестных припасов, как* ***маминька*** *еще тогда* ***обещались*** *прислать сушеных вишен без косточек* [6] *(Also if you send some eatables as in due time mum promised to send some dried seedless cherries).* Moreover this example is interesting because in Russian the noun and verb are not coordinated in number.

Below one can see a part of the table for the verb «иметь» (“to have”).



has_obj4	115	9.4
честь	22	9.14
обыкновение	6	8.8
право	15	7.66
намерение	7	7.54
дар	2	7.21
нужда	5	7.1
возможность	5	6.86
неудовольствие	2	6.54
надежда	5	6.5
успех	2	6.08
удовольствие	3	6.03
цена	2	5.89
характер	2	4.89
вид	2	4.68
<	3	2.87
время	2	1.56

has_obj2	83	8.0
право	8	7.22
возможность	4	7.01
понятие	2	7.01
удовольствие	4	6.92
твердость	2	6.92
власть	2	6.84
талант	2	6.76
значение	2	6.61
терпение	2	6.48
надобность	2	6.42
средство	3	5.59
известие	3	5.29
сила	7	5.14
дух	2	4.58
время	9	4.2

has_subj	69	3.9
власть	5	8.42
талант	2	7.02
цена	2	6.63
недостаток	2	6.18
случай	5	6.17
желание	3	5.93
плетнев	2	5.69
ум	2	5.02
дело	3	3.43
слово	2	3.39
человек	2	2.04

post_prep	142	2.1
над	6	5.02
обо	4	3.55
под	2	2.99
при	3	2.68
ко	3	2.51
об	3	2.31
от	12	2.27
у	6	2.23
для	6	1.94
о	9	1.82
до	4	1.67
на	20	1.29
в	43	1.2
во	3	1.16
к	7	0.94
с	5	0.24

Fig. 2. Example of phrases with the key word «иметь» (“to have”)

The name of each column reflects syntactical relations (has_obj4 – has an object in accusative case, has_obj2 – has an object in genitive case, has_subj – has a subject, post_prep – preposition in postposition). As we can see from the first column the most frequent phrases are «*иметь честь*» (“to have the honour”), «*иметь обыкновение*» (to be in the habit) and «*иметь право*» (“to have right”): *Целуя без-*

ценые ручки ваши ***имею честь*** *быть, с сыновным моим к вам высокопочитанием ваш послушный сын Николай Гоголь Яновский* [7] *(Kissing your beloved hands I have the honour to be with my filial respect to you your obedient son Nikolay Gogol Yanovsky)*; ***Я имею право*** *сказать это, как человек, проведший в борьбе с собой многие годы жизни и лишеньями добившийся до этого права* [8] *(I have right to say it as the man who has spent many years wrestling with himself and owing to hardships has gained this right).*

5 Conclusion

There is a question of corpus volume. For example, we know that different association measures extract different collocations but here one can't see differences between results obtained by a number of statistical measures, it means that collocates will be quite the same. This problem arises from low frequencies of words and phrases.

If we compare data obtained on the corpus of letters of N.V. Gogol to data obtained corpora of literary texts, we can often see considerable differences in collocability that reflect author's characteristic word usage. Thus it can be said that methods described can be effectively used for studying the authors' language and writing authors' dictionaries, for revealing collocability of words in different styles or within the given time period.

References

[1] Kilgariff, A., Rychlý, P., Smrž, P., Tugwell, D. (2004). The Sketch Engine. In: Proceedings of EURALEX-2004, 105-116.

[2] Rychlý, P., Smrž, P. Manatee, Bonito and Word Sketches for Czech. (2004). In Trudy mezhdunarodnoy konferentsii "Korpusnaja lingvistika-2004": Sbornik dokladov. St.-Petersburg, 324-334.

[3] Benko, V. Word Sketches for the Slovak National Corpus. In: Mondilex-2009. Metalanguage and encoding scheme design for digital lexicography (Innovative Solutions for Lexical Entry Design in Slavic Lexicography) Open Workshop (Bratislava, Slovakia, 15th – 16th April 2009)

[4] Gogol, N.V. Polnoye sobraniye sochineniy: [V 14 t.]. (1937–1952). Moscow – St.-Petersburg, 1937–1952. T. X-XIV.

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] Gogol, N.V. Polnoye sobraniye sochineniy: [V 14 t.]. (1937–1952). Moscow – St.-Petersburg, 1937–1952. T. X. Pis'ma, 1820–1835, 41.

[7] Gogol, N.V. Polnoye sobraniye sochineniy: [V 14 t.]. (1937–1952). Moscow – St.-Petersburg, 1937–1952. T. X. Pis'ma, 1820–1835, 31.

[8] Gogol, N.V. Polnoye sobraniye sochineniy: [V 14 t.]. (1937–1952). Moscow – St.-Petersburg, 1937–1952. T. XII. Pis'ma, 1842–1845, 250.

‘New/novelty’ Concept Set Dynamics as a Marker of Lexical and Grammatical Paradigm Evolution for Psychology Sublanguage

Oksana S. Kozak

Kyiv National Linguistic University, Ukraine

Abstract. The paper covers on the theoretical foundations and research data analysis for the specific features of the NEW/NOVELTY concept set dynamics in the psychology sub-language texts. The analysis has been carried out through pointing out and statistically studying the set units on the basis of scientific papers which were published in the 19th and 21st centuries.
Keywords: lexical and grammatical paradigms, a concept set, an invariant, the centre (nucleus), the periphery, a concept diagram.

The statement that language is experiencing continual changes, i.e. that it is in the state of dynamics has become a postulate. It is the process of communication that spots the language dynamics [1, 24]. The written record of communication traces the state of the language system at a certain stage, the stage being performed in definite speech situations. This enables linguists to carry out the comparative analysis of the functional styles on the basis of experimental material where language units get their embodiment in speech, thus providing the chance of taking extra-linguistic factors into consideration. Articles in psychology published in scientific proceedings within two time periods – the second half of the 19th and the beginning of the 21st centuries – served the material for this research. As both the language structure and its functioning are subordinated to statistical rules, they are to be studied with the help of statistical methods [2, 7]. The dynamics of functional semantic and stylistic categories of a certain sub-language can be traced through the changes in lexical and grammatical paradigms. In our opinion, these changes can lead to the changes in the structure of concepts. The above mentioned dynamics is characterized by the system of markers which in our research are represented by the most frequent concept sets. **A concept set** stands for the complex of lexical and grammatical units that serve as the embodiment for one and the same concept in the texture of text.

The study of concept set dynamics is **topical** nowadays as it provides the researcher with the data on the main tendencies and trends in the development of lexical and grammatical paradigm of a certain style sub-language, it also widens the research horizons for compiling, modifying and updating term vocabularies. This paper **aims** at presenting the results of the study on identifying the logic of the NEW/NOVELTY concept set dynamics via figuring out the changes in the set structure in the texts of the two periods. The concept set that has been defined and studied within this research, fits into the plane of lexical and grammatical abstraction and cor-

relates with the concept field through the paradigm of lexical and grammatical units. In their form the elements of the concept set are common root lexemes which stand for one and the same concept and represent the scope of declension range for the notional parts of speech (noun, adjective, pronoun, verb and its non-finite forms) [3; 4].

The task of this research has been to define the most frequent units of the NEW/NOVELTY set, to identify absolute and relative set units frequency for each of the two periods under research, as well to carry out the analysis of set frequencies according to their dispersion, range, statistical homogeneity, number of set units.

We assume that the set of the most frequent units of the concept set correlates with the specific features of the functional semantics and style categories (FSSC) [5, 139] and their main traits for a certain period.

We state that the dynamics of lexical and grammatical paradigms of the psychology sub-language is the process of its functional and stylistic development, improvement, thus it results in broadening and specifying the properties of the units of the above mentioned categories for the best service of the communication aims in scientific sphere for a certain period.

We have analyzed the texts of Psychology scientific papers with the overall word usage number of 70 400 (35000 and 35400 word usage cases for the first and the second period respectively). Thematically related texts for the two periods have been analyzed. The analysis resulted in defining the most frequent NEW/NOVELTY concept units for both of the compared text corpora. The total number of units indicates the theme breadth of the articles under study. For Period 1 (the second half of the 19th century) overall units frequency per corpus is 17, for Period 2 (the beginning of the 21st century) it is 38.

The absolute frequency of concept units has been calculated followed by the analysis of their relative frequencies for the two periods according to:

a) frequency dispersion (the number of texts for each of the periods under research in which the concept units are represented);

b) frequency range (diapason):

- low (0.01 – 0.09%);
- lower than the medium (0.10 – 0.19%);
- medium (0.20 – 0.29%);
- higher than the medium (0.30 – 0.39%);
- high (0.40% and more);

c) frequency homogeneity (statistically homogeneous/ non-homogeneous frequencies);
d) the quantity of the concept set units for a certain period.

Here below is the table of comparison for the units frequency of the NEW/NOVELTY concept set for the two periods:

	PS 1	p = 0.05%
№	**Unit**	**Frequency**
1	neo-	0
2	**new**	13
3	newly	0
4	news	0
5	newsy	0
6	renew	0
7	renewable	0
8	renewal	0
9	renewed	3
10	novel	0
11	novelty	1
12	novelties	0
	∑	**n=17**

	PS 2	p = 0.1%
№	**Unit**	**Frequency**
1	neo-	4
2	**new**	22
3	newly	0
4	news	9
5	newsy	0
6	renew	0
7	renewable	0
8	renewal	0
9	renewed	0
10	novel	1
11	novelty	1
12	novelties	1
	∑	**n=38**

Table 1. Correlation between the units quantity for the NEW/NOVELTY Concept Set (Periods 1, 2)

where PS 1 and PS 2 stand for the first and the second periods of papers on Psychology, m and n – relative and absolute set frequencies respectively, the unit in bold stands for the core of a concept set of a certain period [6, 107].

Frequency dispersion analysis has provided the possibility to find out that the dispersion of the concept set units is the same. Should the set be represented in the texts of one of the two periods only, it could be the indicator of increase/decrease in evincing interest for a certain psychological issue.

Frequency range (diapason) analysis has enabled not just to trace the vector of the NEW/NOVELTY, but has also prompted the idea of building concept diagrams for both separate articles [5] and for the psychology sub-language in general for each of the two periods under analysis in accordance with the categories and the core/periphery frequencies.

Frequency homogeneity analysis has made it possible to estimate the statistical frequency homogeneity/absence of homogeneity for the concept set units of each of the two periods.

The quantity of the concept set units analysis has enabled to trace the dynamic changes of lexical and grammatical paradigms within the concept set (the set centre shift from one unit onto another; increase/decrease in the relevant frequency of the concept set; the change in lexical and grammatical unit status.

According to the frequency dispersion the NEW/NOVELTY conceptual set is a prevailing one with a non-significant difference in frequency. According to the fre-

quency range (diapason) the set belongs to different groups for the two different periods: low frequency (0.01 – 0.09%) group (for the first period); lower than the medium frequency (0.10 – 0.19%) group. According to the frequency homogeneity the NEW/NOVELTY is a set with statistically non-homogeneous frequencies. There is also a difference in the qualitative set composition, as well as in its lexical and grammatical status. Although the data from Table 1 do not show the significant shifting of the concept core (new 13 → new 22), just show its slight movement in the direction of news 0 → news 9, still the increase in the core units frequency accompanied by an inconsiderable change in the quantity of units variants in the set periphery for both of the corpora (novelty 1 → novel/novelty/novelties 3), inspires us to make a conclusion about the increase in textuality loading on the NEW/NOVELTY concept in the texts of the Period 2. Moreover, a significant change in the qualitative set composition in the set periphery has been spotted (neo- 0 → neo- 4; renewed 3 → renewed 0). The results of the analysis are given in Table 2 below:

Period	Dispersion	Range (Diapason)	Homogeneity	**Quantity of Units**	P%
Period 1 **(the second half of the 19th century)**	100%	Low (0.01 – 0.09%)	-	3	0.03-0.06
Period 2 **(the beginning of the 21st century)**	100%	Lower than the medium (0.10 – 0.19%)	-	6	0.05-0.16

Table 2. Unit Frequency for the NEW/NOVELTY Concept Set (Periods 1, 2)

The analysis has enabled to arrive at the conclusion that the concept under research has a significantly different weight in the texts of the two periods, as well as to construct a concept diagram for the NEW/NOVELTY set as based on the data received.

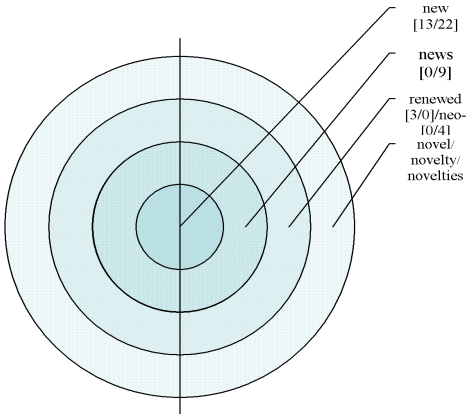


Chart 1. NEW/NOVELTY **Concept Diagram** (Periods 1, 2)

There is no doubt that the results of the research prove the need for a further research at the level of each of the of the functional semantics and style categories of the psychology sub-language [5], which will make it possible to build the above specified diagrams.

References

[1] Левицький, А. Э. (2002). Особенности динамики лексико-грамматической системы современного английского языка // Мова і культура (Науковий щорічний журнал). – К.: Видавничий Дім Дмитра Бураго,. – Випуск 5. – Т III. Ч.: Національні мови і культури в їх специфіці і взаємодії. – С.23–28.

[2] Перебийніс В. І. (2001). Статистичні методи для лінгвістів: Навчальний посібник. – Вінниця: Нова книга. – 168с.

[3] (2003). Longman Dictionary of Contemporary English. – Harlow, England: Pearson Education Limited. – 1950 p.

[4] Hornby A. S. (2006). Oxford Advanced Learner’s Dictionary. – Oxford: Oxford University Press.– 1715 p.

[5] Козак О. С. (2006). Мовленнєва системність та стилістичні категорії підмов філософії та психології // Проблеми семантики, прагматики та когнітивної лінгвістики: Зб. наук. пр. Вип. 10 / Київ. нац. ун-т ім. Т. Шевченка; Відп. ред. Н. М. Корбозерова. – К.– С. 139–143.

[6] (2007). Computer Treatment of Slavic and East European Languages // Fourth International Seminar: Bratislava, Slovakia, 25 – 27 October 2007. – Tribun: Bratislava. – P. 104 – 108.

Methodological Foundations for Contrastive Model of Verb Valence

Ružena Kozmová

University of SS. Cyril and Methodius in Trnava, Slovakia

Abstract. Subject of this contribution is to present contrast model of the verb valence and methodology sources, which are used in the process of model building.
In the comparison of the verb valence in Slovak and German, we can see general differences (theoretic basis of the valence, theory of the verb intention, terminology, structure of the sentence) and specific differences (minor differences in the valence, some differences in the actionality and major differences in the variations of verb meaning).
In the summary we present sentence models of verb *sagen*, to which corresponds Slovak equivalents of perfective verb *povedat'* and imperfective verb *hovorit'*.

Prologue

Subject of this contribution is presentation of contrastive model of verb valence. We want to concentrate mainly on the methodological foundations that apply in the model construction. In the confrontation of verb valence in Slovak language and German language appears general differences (theoretical basis of the valence, theory of the verb intention, characteristic of verb and verbal situation, resp. predicates, terminology, model structure) and specific (minor differences in the valence in general, absence of the aspect category in German and major differences in so called semantic verb variations).

Attention also deserves clarification of contrastive model of verb valence concept, which we understand in the sense, that in the model construction of given verb, both languages are target as well as source. Specifically it means description of sentence model of verbal lexeme in German language with presentation of its equivalents in Slovak language. And vice versa model of verbal lexeme in Slovak language and its equivalents in German language. This procedure is actually contrastive – confrontational and its result is always two models of verbal lexeme. Construction of contrastive-confrontational model has from the mentioned reason specific character; therefore it will differ from both German as well as Slovak model of verb valence used in the existing valence dictionaries[1].

[1] We have in mind valence models of German verbs beginning with the oldest as is G. Helbig/ Schenkel, W. (1983): Wörterbuch zur Valenz und Distribution deutscher Verben. Tübingen: Max Niemeyer and ending with the most recent: H. Schumacher/ J. Kubczak/ R. Schmidt, V. de Ruiter (2004): VALBU – Valenzwörterbuch deutscher Verben. (Studien zur deutschen Sprache 31). Tübingen: Narr. For Slovak language I quote so far the single valencedictionary: J. Nižníková/

1 To theory of valence and intention

Concerning the limitation of space we have to be satisfied with only the outline of the theory of valence development in the German linguistics, that has been determined not only by European[2], but also by American linguistics. Our goal is to demonstrate, that two on the outside diverse theories: theory of valence and theory of verb intention stands in the complementary relation and can create optimal theoretical base for the characteristics of sentence formula, or sentence model of verb.

Theory of valence, established by Tesnière[3], found great response in German linguistics and relatively soon in the frame of its development two streams were detached according to dominance of linguistics levels. This way, in the beginning of seventies and eighties a group of linguist was singled out, which was researching valence on the morphosyntactic base[4], mainly under the leadership of G. Helbig (G. Helbig-M.D. Stepanova, G. HelbigW. / Schenkel etc.) and so called semantic group, that was represented mainly by K. Welke, K. Bondzio, K.-H. Sommerfeldt. G. Helbig later[5]reassessed theoretical foundations of valence in previous period and outlined six stages complex model of verb valence – model emerging from syntax – because after all, it is the supplementing of sentence – however, its basis is semantic layer and by way of semantic markers it is possible to specify more accurate sentence model and its participants. Sentence model is elementary problem in the composition of sentence also according to Ružička[6] (1968), because sentence and sentence models are important instruments of each statement. It is therefore desirable to look for these criteria in „features, thus semantic markers of verb as word class." For Ružička the most important characteristics of verb are valence and intention of verbal situation, which was elaborated by Eugen Pauliny[7]. Such an understanding is in Slavonic[8] typical for many

Sokolová, M. (1998): Valenčný slovník slovenských slovies. FF Prešov.

[2] Let's recall from older period J. A. Lehmann and his work„Allgemeiner Mechanismus des Periodenbaues" z r. 1833. According to Lehmann, verbum finitum is important from formal and material aspect, however, subject is from formal aspect irrelevant, but from material aspect relevant. Model construction works following the principle of dependency: „Subjekt sei „Nichtnotwendiger oder Nebenbestandteil des ersten Ranges", other sentence membersare members of second level, hence „Dependenzen".

[3] Concept of *valence* was originally adopted from chemistry and it was taken into linguistics independently by S. D. Kacneľson (1948) and de Groot (1949). Foundation of the theory was formulated by L. Tesnière (1959): „Éléments de syntaxe strusturale".

[4] Compare works of G. Helbiga z r. 1983, 1987.

[5] G. Helbig (1992): Probleme der Valenz- und Kasustheorie. Tübingen: Max Niemeyer.

[6] J. Ružička (1968): In : Jazykovedný časopis.

[7] E. Pauliny (1943): Štruktúra slovenského slovesa. Bratislava: VEDA.

[8] Ľ. Reháček (1966): Sémantika a syntax infinitivu v současném polském spisovném jazyce. Concept of intention is understood by Reháček as function of lexical semantics of expression on the contrary to valence, which for him represents only grammatical ability to bind other grammatical forms.

linguists until these days. I assume, that the relation among valence and intention was expressed most succinctly by Daneš [9] thereby, that he revealed relation of polysemy among valence and intention. Within components of valence and intention structure dominates asymmetric dualism of form and meaning, hence polysemy. The same valence structure is the base for multiple intention structures. So if we understand valence as a complex grammatical (not only morpho-syntactic phenomenon), than intention is the part of valence. That valence, which is also in the spirit of Helbig's six stages model[10] capable to include all language layers, beginning with morphological and ending with pragmatical[11] layer. In our model, however, we want to follow also the most recent results, achieved in the area of semantic verb research, verbal situation and semantic markers arguments especially in works of V. Ágel[12] and K. Welke[13].

2 To the composition of contrastive – confrontational model:

Valence model of German verb in its latest form[14] connects relatively optimal morpho-syntactic and semantic elements of the valence. Valence model of Slovak verbs has mainly lexical-semantic character. It is not the objective of this paper to point to positives or negatives of both valence dictionaries. To the structure of valence dictionary of Slovak verbs, Jarmila Panevová[15] gave her opinion and she herself suggested functional-generative description of verb valence[16]. In connection to the model we will in accordance with German linguistics[17] distinguish concept of sentence formula and sentence model.

Sentence formulas are abstract categories, whose amendments can be realized with various participants. Sentence model is concrete, referring to particular sentence structure. Even in this regard there are many differences among compared languages.In this way, for example, it is possible in one language to change the subject to the lateral sentence, in other language; however, the equivalent of verb does not allow it. The other fact also needs to be stated and that is, that sometimes does not exist accurate, by convention given translation, equivalents of given verb, actually that for given verb of initial language exists several equivalents in target language[18].

---

[9] F. Daneš / M. Grepl/Z. Hlavsa (1987: 34): Mluvnice češtiny. Praha: Academia.

[10] G. Helbig (1992): Probleme der Valenz- und Kasustheorie. Tübingen: Max Niemeyer.

[11] We understand the relation of systemic linguistics and pragmatics as a mutually overlapping, see also P. Ernst (2004): Einfúhrung in die germanistische Sprachwissenschaft. Wien.

[12] V. Ágel (2000): Valenztheorie. Tübingen: Günther Narr.

[13] K. Welke (2005): Deutsche Syntax funktional. Tübingen: Niemeyer.

[14] H. Schumacher et al. (2004): VALBU Valenzwörterbuch deutscher Verben.

[15] J. Panevová (2005): Sloveso: centrum vety, valence: centrálny pojem syntaxe. In: Aktuálne otázky súčasnej syntaxe, p. 73-74. Bratislava: SAV VEDA.

[16] J. Panevová 1974, 1975, 1980, 1994.

[17] See especially U. Engel (2004): Deutsche Grammatik. Neubearbeitung. München: Iudicium.

In both languages verb has also prepositions with the contrary meaning or German has alternative prepositions (erzählen: von, über) which Slovak does not have. Beside of that, in Slovak are profusely used so called disposal constructions that present German language does not know. In both languages exists verbs, of which it is possible under certain circumstances to leave out Esit/Dsit. In such cases is position of stipulation known or comprehended in the frame of context. In most cases, however, Esit/Dsit realize: goes mainly for verbs: stattfinden/ take place, leben/live, wohnen/live in, sitzen/sit etc.

These and other differences does not allow to create single model of verb in both languages also considering the fact, that there is no metalanguage, which would provide objective and complex tertium comparationis. I assume, that also for this reason most of bilingual valence dictionaries limit itself to use German as initial language and in mother tongue are presented standard equivalents[19]of respective German sentence models.

2.1 Sphere of terminology

In connection to theoretical basis of valence and intention of verb as its part, it is possible to determine also unified terminology for both languages. If we will come out from the premise, that both German and Slovak linguistics consider as the base of sentence the sentence formula, resp. particular sentence model, symbols of sentence complements can be used for both languages. I decided for symbols listed in the latest work of U. Engl[20], therefore we will distinguish among complements subjective: Esubj/Dsubj,objective namely in accusative: Eakk/Dakk, genitive: Egen/Dgen, dative: Edat/Ddat, in prepositional case Epräp/Dprep, local: Esit/Dsit, directive: Edir/Ddir or directional, predicative: Epräd/Dpred and Eexp/Dexp a Everb/Dverb.

In case of adverbial complement, which are being under certain conditions mandatory, e.g. causative, temporal and modal we will use symbol Eadv/Dadv. Instead of term complement it could be possible to use term complements and supplements. I do not like this alternative for two reasons. First from didactic reasons term complement = Ergänzung is commonly used in German language schoolbooks not only on secondary level schools[21], but already also on the elementary schools and second even from linguistic point of view this term is transparent. Subject and object are sentence components not only in the frame of so called traditional syntax, but they have to be supplemented in the sentence so that it is meaningful, and so are complements becoming subjective or objective comprehensible and usable no matter what is the syntactic

[18] R. Kozmová (2007): Zum Problem des kontrastiven Valenzmodells am Beispiel der be-Verben. Wrocław. About to be published.

[19] As an exception can be considered sentence formulas listed in German-polish grammar: U. Engel / D. Rytel et al (1999): Deutsch-polnische kontrastive Grammatik. Heidelberg: J. Groos.

[20] U. Engel (2004): Deutsche Grammatik. Neubearbeitung. München: Iudicium.

[21] G. Neuner et al.: Deutsch aktiv, Deutsch aktiv neu.

theory, moreover in Slovak linguistics[22], term complement is used for nominal predicative complements.

2.2 Categorical meaning

Categorical significance entitles semantic minimum of context components. The minimum requirements for these components are known[23], because long time ago[24], before J. Katz a J. Fodor formulated inventory of semantic markers in their work, they were used in dictionaries and in parallel in many others lexicographic works. There is a well known fact, that the choice of verb conditions selection of substantive, therefore it is logical, that interest for semantic characteristics of verbs and substantives was raising together with theoretical development of valence in Anglo-Saxon, as well as German and Slavonic linguistics.

Concerning analyzed languages, it is necessary to mention works of G. Helbig, who coming out mainly from theory of deep cases of Ch. Filmore, elaborated relatively in most detailed scope of semantic cases[25] in German language. Despite of certain differences in German and Slovak literature, might be used inventory of semantic markers of substantives defined by Ulrich Engl[26], namely in the widest sense of the word, because we identify with the above mentioned author in that hierarchy of semantic markers, videlicet not only substantives, is relatively difficult. The more precise we want to define semantic markers, the more we hit on quantity of not always secondary problems. On this account, but mainly because, as we deal here with two different languages, we will not continue as Sokolová in VSS[27].

2.2.1 Interdependency of semantic markers of verb and substantive

The choice of functional-semantic method in the analyzing of grammatical phenomenon was in many cases shown as optimal and that is also in the case of the verb valence. Its ability to bind to itself certain complements is nothing else than ability of one syntactical unit, by way of semantic markers, form meaningful syntagma on the basis of compatible markers of other syntactical unit of the same or other word class.

---

[22] See M. Sokolová 1998: 46.

[23] J. A. Katz / J. J. Fodor (1963): The Struture of a Semantics Theory.In: Language 39. p.170-210.

[24] W. Porzig (1943): Wesenhafte Bedeutungsbeziehungen. In: Beiträge zur Geschichte der deuschen Sprache und Literatur. p.70-97.

[25] In his most recent grammarG. Helbig presents 20 semantic functions: G. Helbig (2002: 468): Deutsche Grammatik. Ein Handbuch für Ausländer. Berlin / München / Wien / Zürich / New York: Langenscheidt

[26] U. Engel (2004: 188): Deutsche Grammatik. Neubearbeitung. Múnchen: Iudicium

[27] J. Nižníková /M. Sokolová (1998): Valenčný slovník slovenských slovies. Prešov: Slovacontact.

We will come out from the fact, that verb has inherence meaning, by means of which it binds with other word classes on the base of their categorical meanings[28]. Following it is combinational characteristics of verb, because it allows binding with participants and since they enter into relations, it needs to talk also about relational meaning. All these characteristics of verb determine relation of interdependence of semantic markers of individual participants and semantic markers of verb, and therefore they participate on the creation of sentence formula, which has its specific utilized shape in the sentence model. Even the smallest components, semems are expression of asymmetric form's dualism and function, because only one sentence formula can have various semantic partners – participants. It does not have to be always only agens (Peter cooks a soup), as well as it can be a carrier of process (The soup is being cooked) or carrier of state (The soup is cooked). Even agens can have by means of diversely ordered semantic markers of verb and substantive various modifications. German and Slovak language as accusative languages have more achievements and accomplishment verbs as verbs representing process. Therefore it is logical, that achievements verbs are becoming process verbs apart that the syntactical structure would change. In the spirit of prototypes theory Welke[29] explains it in following way:

Prototypic situation has human agens [+hum] as left participant, which deliberately (with own will) in the spirit of intention [+intent] on the base of own activity [+control] evokes situation, maintains it, suspense and is at the same time responsible for it [+responsible], thus for its ending, preservation or suspension. Prototypic effect according to Welke resides in that in case of loss of some marker, e.g. adjacent remains, therefore existence of agens is only weakened (metaphorically), but agens even though remains, as it is confirmed by sentence 1, 2.

(1) Petra cooks a good soup.
(2) Our kitchen cooks a good soup.

2.2.2 Verb, verb situation, predicates and actionality in language

In connection to verb situation, resp. with actionality we need to keep in mind great differences in both analyzed languages. Slovak language is aspectual; aspect is in this language grammatical category that influences the whole verbal system. German language does not know such a category, and so duration, or one of phases of verb situation is usually created analytically. Vague semantic even of so called true perfective verbs is more or less evidently demonstrated in equivalents of German lexeme[30]. Lemma raten/radiť usually represents imperfect verb situation, but on the examples 3a, b we can see, that the same sentence can be expressed as perfective as

[28] For questions in semantic discipline see also works of J. Kačala (1989): Sloveso a sémantická štruktúra vety, M. Sokolová: (1993): Sémantika slovesa a slovesný rod. Bratislava: VEDA.

[29] K. Welke (2005: 216): Deutsche Syntax funktional. Tübingen: Niemeyer

well as imperfective aspect. And one more example:Verb *merken* (4) as well as *bemerken* (5) has always one equivalent, and that is perfective verb *zbadať.*

(3) Ich rate dem Kunden, ein anderes Geschäft aufzusuchen.
(3a) Poradím zákazníkovi, aby vyhľadal iný obchod.
(3b) Radím zákazníkovi, aby vyhľadal iný obchod.

(4) Inzwischen hat auch er gemerkt, dass es so nicht geht.
(4a) Medzitým zbadal, (uvedomil si), že to takto nejde.
(5) Das habe ich [gar nicht] bemerkt.
(5a) To som [vôbec] nezbadal (nespozoroval).

In the summary we present sentence models of verb *sagen*, to which corresponds Slovak equivalents perfective verb *povedať* and imperfective verb *hovoriť*, even though to Slovak lexeme *hovoriť* corresponds German lexeme *sprechen.* This fact only confirms our previous assertion about vagueness of German verb.

Sentence models: sagen/povedať(say)
a) meaning of verbs sagen, povedať

Meaning	**Example sentences**	**Equivalent in Slovak**
äußern, aussprechen	Er hat mit anderen Worten das Gleiche wie der Vorredner gesagt.	povedať
benutzen, gebrauchen (ein Wort)	Sagt man „über jemanden“ oder „von jemandem“ sprechen.	**hovoriť/sagen**
bezeichnen	Der Fachmann sagt zum Schraubenzieher Schraubendreher.	(ne)povedať
bei jm. etwas bewirken	Eine innere Stimme sagt ihm, dass alles gut sein wird.	(na)povedať, **našepkať**
informieren (über eine E.)	Das haben Sie schön gesagt.	povedať
vorschreiben, befehlen:	Du hast mir gar nichts zu sagen!	**hovoriť**
vermitteln, ausdrücken, aussagen, berichten	Niemand kann wirklich Verläßliches über die Höhe der Investionen sagen.	povedať
bemerken, feststellen	Das hat er tatsächlich zu mir gesagt!	povedať
meinen, behaupten, M. ver-	Was sagt denn dein Vater	**hovoriť**

[30] For aspect questions in connection with Slovak equivalents of German lexeme look into work of:R.Kozmová: Zum Problem des kontrastiven Valenzmodells am Beispiel der be-Verben. Paper on international conference: „Germanistische Linguistik extra muros“ - Inspirationen, Aufgaben, Aufforderungen.Wroclav 2008.

treten	dazu , dass du schon rauchst?	
argumentieren	Dagegen ist nichts zu sagen.	povedať, **namietať**
annehmen, glauben :	Ich sage, es gibt heute noch Regen.	**myslieť si, domnievať sa**
formulieren	Das kann man einfacher sagen	povedať
zeigen	Der Film sagt mit eindrucksvollen Bildern: Krieg bedeutet Tod.	ukazovať
sich denken, sich überlegen	Da hab ich mir gesagt: Was dem einen recht ist, ist dem andern billig.	**(po)myslieť, uvedomiť si**
schweigen	Er sagte kein Wort.	**mlčať**
beurteilen	Ich kann über dieses Bild- nichts Positives sagen.	povedať, **posúdiť**
zum Inhalt haben	Das Gesetz sagt eindeutig, dass ...	**hovoriť**
als Schluss zulassen, besagen, heißen	Das allein sagt noch nicht viel.	**(ne)hovoriť**
(Worte, Äußerungen) an jmdn. richten	Er sagte ein paar aufmunternde Worte und verschwand.	povedať
erzählen mit Dat	Ich habe mir sagen lassen.	**dať si porozprávať**
sich entfremden	Sie haben sich nichts mehr zu sagen.	povedať

Povedať:

Meaning of the verb *povedať*	**Example sentences**	**Equivalent in German**
vyjadriť slovami, rečou:	Povedal, že sa nemáme znepokojovať	sagen
vysloviť názor, zaujať stanovisko	Čo mu na to povedia?	sagen
predniesť	Anička povedala básničku.	sagen
prezradiť	Poviem ti veľkú novinu.	sagen
určiť, stanoviť	Predavač povedal cenu.	sagen
povedať si, dohovoriť sa, dohodnúť sa	Povedali si, že sa stretnú pred kinom.	(sich) sagen/sich verabreden
pomenovať, nazvať	Odborník nepovie šrobovák ale skrutkovač.	sagen
sľúbiť	Povedal, že príde.	sagen
obohatiť poznatkami, vysvetliť:	Táto teória nám povedala veľa.	(be)sagen

hovoriť pre seba, opakovať si	Povedala si, čo sa naučilla.	(sich) zu sich sagen
žalovať,obviniť niekoho z niečoho	Povedal to na mňa otcovi.	**petzen**
rozhodnúť sa	Povedz posledné slovo!	sagen
uznať	Daj si povedať a poď s nami!	sagen
niečo namáhavo presadzovať	Povedz pravdu, prebiješ si hlavu.	
spôsob formulácie	To ste pekne povedali.	sagen

b) generally formula of verb sagen

Verb formula (VF): Esubj, Eakk/Epräp = von + Dat, über + Akk, zu + Dat, (Edat), Eadv

Semantic marker (SM): sagen/povedať

Esubj = [+control], [+intent], [+reversible], [+hum], [+konkr], [+immat], [intell]
Eakk= [+sachv], [mas/indefinit], [+immat], [+intell]
Edat = [+hum], [+inst]
Epräp = [+hum], [+sachverh]
Eadv/mod = {sonst], [+qual}
NS = dass-S, w-S, ob-S, uNS

c) sentence models:
E = D = complement/doplnenie
NS = VV = subordinative sentence

sentence model 1 Esubj, NS /Esubj, (Korr), NS Esubj, NS/HS	Ich sage, es gibt heute noch Regen. *Myslím si, že dnes bude ešte pršať.*	annehmen, glauben, meinen, behaupten, M. vertreten, ukázať, zum Inhalt haben

sentence model 2 Esubj, Eakk	Er hat [mit anderen Worten] das Gleiche wie der Vorredner gesagt.*Povedal [inými slovami] to isté, čo jeho predrečník.*	äußern, aussprechen, benutzen, gebrauchen (ein Wort), formulieren, schweigen (nicht sagen), Worte, Äußerungen an jemanden richten

sentence model 3 Esubj, Edat	Da hab ich mir gesagt: Was dem einen recht ist, ist dem andern billig. *A tak som si povedal: Čo jednému vyhovuje, to druhému vadí.*	sich denken, sich überlegen

sentence model 4 Esubj, Epräp	Dagegen ist nichts zu sagen. *Voči tomu nemožno nič namietať.*	argumentieren, (nicht) einwenden

sentence model 5 Esubj, Eakk	Das allein sagt noch nicht viel. *To samo osebe ešte nič neznamená*	als Schluss zulassen, besagen, heißen

sentence model 6 Esubj, Eakk, Epräp/(Epräp)	Der Fachmann sagt zum Schraubenzieher Schraubendreher. *Odborník nepovie šróbovák, ale skrutkovač.*	vermitteln, ausdrücken, aussagen, berichten, bemerken, feststellen

sentencemodel 7 Esubj, Eakk, Edat	Du hast mir gar nichts zu sagen! *Ty mi nemáščo hovoriť (rozkazovať).*	vorschreiben, befehlen

sentence model 8 Esubj, Edat, NS	Eine innere Stimme sagt ihm, dass alles gut sein wird. *Vnútorný hlas mu našepkáva, že všetko bude dobré.*	bei jemandem etwas bewirken.

sentence model 9 Esubj, Eakk, Eadv	Das haben Sie [schön] gesagt. *To ste pekne povedali*	informieren(über eine Eigenschaft)

sentence model 10 Esubj, Edat, Everb	Ich habe mir sagen lassen. *Dal som si porozprávať.*	erzählen mit Dat

c) generally formula *povedať*

Verb formula (VF): Esubj, Eakk/Epräp =na + Akk , Edat/(Edat), Eadv/mod

Semantic marker (SM): sagen = povedať

VV = že, aby, ako, aký, kedy, či, čo, prečo, s kým

sentence model 1 Dsubj, Dakk	Anička povedala básničku. *Anička trug ein Gedicht vor.*	predniesť, povedať, stanoviť, rozhodnúť sa, niečo namáhavopresadzovať, pomenovať, nazvať

sentence model 2 Dsubj, Ddat	Čo mu na to povedia? *Was sagen Sie ihm dazu?*	vysloviť názor, zaujať stanovisko

sentence model 3 Dsubj, Dprep	Daj si povedať a poď s nami! *Lass dich überzeugen und komm mit!*	uznať

sentence model 4 Dsubj, VV	Povedal, že sa nemáme znepokojovať *Er sagte, wir sollten uns nicht beunruhigen.*	vyjadriť slovami, rečou

sentence model 5 Dsubj, Dexp	Táto teória nám povedala veľa. *Diese Theorie sagte uns sehr viel.*	obohatiť poznatkami, vysvetliť

sentence model 6 Dsubj, Dakk, Ddat	Poviem ti veľkú novinu. *Ich sage dir eine große Neuheit.*	prezradiť

sentence model 7 Dsubj, Ddat, Dprep	Povedal to na mňa otcovi. *Er petzte bei dem Vater über mich.*	žalovať, obviniť niekoho z niečoho

sentence model 8 Dsubj, Dakk, Dadv	To ste pekne povedali. *Das haben Sie schön gesagt.*	spôsob formulácie

sentence model 10 Dsubj, Ddat, VV	Povedali si, že sa stretnú pred kinom. *Sie haben sich verabredet, dass sie sich vor dem Kino* treffen.	povedať si, dohovoriť sa, dohodnúť sa

References

Ágel, Vilmos (2000): Valenztheorie. Tübingen: Narr

Daneš, František / Zdenek Hlavsa(1981): Větné vzorce v češtině. Praha: VEDA

Dolník, Juraj (1990): Lexikálna sémantika: Bratislava: FF UK.

Drosdowski, Günther et al. (1989): Deutsches Universalwörterbuch A. Z. Mannheim/Wien/Zürich:DUDEN © Duden - Deutsches Universalwörterbuch, 5. Aufl. Mannheim 2003 [CD-ROM]

Ďurčo, Peter (2009): Modellbeschreibung für Kollokationen mit Basiswort Verb. S.67-79. In: Deutsche Sprache in der Slowakei.Festschrift fúr Prof. Dr. Ilpo Tappani Piirainen zum 65.Geburtstag. Internationale Fachtagung Piešťany 13.-15. Juni 2007.

Engel, Ulrich (1976): Liste deutscher Verben mit ihren Bedeutungsbeschreibungen. In: Engel, Ulrich /Helmut Schumacher u.a.:Kleines Valenzlexikon deutscher Verben. (Forschungsberichte des Instituts für deutsche Sprache 31). Tübingen: Narr. Engel, Ulrich (1988): Deutsche Grammatik. Heidelberg: J. Groos.

Engel, U./D.Rytel et al. (1999): Deutsch- polnische kontrastive Grammatik. Bd. 1. Heidelberg: J. Groos.

Engel, Ulrich (2004): Deutsche Grammatik. Neuberbeitung. München: Iudicium.

Engel, Ulrich/Meike Meliss. Hrsg. (2004): Dependenz, Valenz und Wortstellung. München: IUDICIUM.

Helbig, Gerhard/M. D. Stepanova (1978): Wortarten und das Problem der Valenz in der deutschen Gegenwartssprache. Leipzig: VEB Bibliographisches Institut.

Helbig, Gerhard /W. Schenkel, Wolfgang (1983): Wörterbuch zur Valenz und Distribution deutscher Verben.Tübingen: Max Niemeyer

Helbig, Gerhard (1992): Probleme der Valenz- und Kasustheorie.

Helbig, Gerhard, Joachim Buscha (2002): Grammatik der deutschen Sprache. Ein Handbuch für Ausländer. Berlin /München/ Wien/ Zürich/ New York: Langenscheidt.

Kačala, Ján (1989): Sloveso a sémantická štruktúra vety. Bratislava: VEDA.

Kačala, Ján (1998): Syntaktický systém jazyka: Pezinok: Formát.

Kačala, Ján et al. (2003): Krátky slovník slovenského jazyka. Bratislava: VEDA.

Katz, Jerry./J. A. Fodor (1963). The structure of a Semantics Theory. In: Language 39. S.170-210.

Lakoff, George/Mark Johnson (1999): Philosophy in the Flesh. In: Computational Linguistics, vo. 25.

Nižníková, Jolana/Sokolová, Miloslava(1998): Valenčný slovník slovenských slovies. FF Prešov.

Panevová, Jarmila. (2005): Sloveso:centrum vety, valence: centrálny pojem syntaxe. 73-78. In: Aktuálne otázky súčasnej syntaxe.Bratislava: VEDA SAV.

Pauliny, Eugen (1943): Štruktúra slovenského slovesa. Bratislava: VEDA.

Peciar, Štefan et al. (1959-1968): Slovník slovenského jazyka. Bratislava: VEDA

Schumacher, Helmut./Jacqueline Kubczak/Renate Schmidt, Vera de Ruiter (2004): VALBU – Valenzwörterbuch deutscher Verben. (Studien zur deutschen Sprache 31). Tübingen: Narr.

Schreiber, Herbert/ Sommerfeldt, Karl-Ernst/Starke, Günter (1993): Deutsche Substantive. Wortfelder für den Sprachunterricht: Berlin/München u.a: Langenscheidt – Verlag Enzyklopädie

Sokolová, Miloslava (1993): Sémantika slovesa a slovesný rod. Bratislava: VEDA

Storrer, Angelika (1992): Verbvalenz. Theoretische und methodische Grundlagen ihrer Beschreibung in Grammatikographie und Lexikograpie. Tübingen: Niemeyer.

Tibenská, Eva (1998): Objektový participant sémantickej štruktúry vety. In: Slovenská reč, roč. 63, č. 4, s. 198-209.

Tibenská, Eva (2005): Význam a forma v syntaxi. S. 28-36. In: Aktuálne otázky súčasnej syntaxe. Bratislava: SAV.

Tomášiková, Slavomíra (2008): Ekvivalenty slovenských lokálnych predložiek v štandardnej a rakúskej nemčine. Acta facultatic philosophicae universitatis. Prešoviensis. Prešov: FF. Schumacher, Helmut (1986): Verben in Feldern.

Valenzwörterbuch zur Syntax und Semantik deutscher Verben. (=Schriften des Instituts für deutsche Sprache 1). Berlin/New York: de Gruyter.

Welke, Klaus (2005): Deutsche Syntax funktional, Tübingen: Niemeyer.

Dictionary of Štúr's Slovak

Ľubomír Kralčák

Faculty of Arts, Constantine the Philosopher University in Nitra, Slovakia

Abstract. Štúr's Slovak language represents a special landmark signifying formation of modern literary Slovak language in the history of literary language of Slovaks. Specific features of Štúr's language display at every language level, whereas vocabulary belongs to those levels, which embody the most numerous representations of the particularities like this.

Dictionary of Štúr's Slovak Language is a project that shall capture the vocabulary of Štúr's period as a precious cultural heritage at maximum extent. Hence, it will be necessary to answer the question of a source dictionary base, i.e. list of source texts, according to maximum model, in vertical restriction (Štúr's texts from the broadest period) and in horizontal restriction (all kinds of texts from the given period). The project assumes, after setting the source base, the transfer of the texts into the electronic form, and on this base and by means of verified software, the creating of lexical data base and its subsequent lexicographical elaboration.

Štúr's Slovak represents a peculiar landmark in the language history of Slovaks signifying the beginning of modern standard Slovak. Yet in case of Štúr's Slovak it is a concept establishing the structure of modern standard Slovak. Comparing it to the present-day Slovak, there are considerable differences and peculiarities which give this language formation a unique form. Specific features of Štúr's Slovak show in all the language levels whereby the vocabulary belongs to those subsystems which show the most numerous representation of such peculiarities. Formation of standard Slovak during Štúr's period represents simultaneously incredibly vivid, dynamic phase of formation and that not only from grammatical and syntactical view but also from the view of the range of functional styles. Lexis has a special status in this context thusly representing the mentioned unique dynamics most clearly.

1 Štúr's lexis

This lexis reflects formation processes in a broader sense: it is not only the formation of the reformulated standard language alone, drafted according to the medieval language structure but also the current social development. Its indicators were new knowledge, products and social phenomena. Extra language innovations were implemented into the language as an adequate accomplishment of communication needs in the form of vocabulary innovations.

This process can be first seen in the formation of the so-called cultural lexis which became the main area of the vocabulary enrichment in the mentioned period. Cultural lexis is formed by nomination of general abstract concepts, social, cultural, civilizational

phenomena and technical terms. The authors of the Štúr's lexical norm used all the ways of natural vocabulary enrichment, that is it was about enlarging the meaning of domestic words, formation of new namings, transfer of foreign words and translation of foreign words (comp. Horecký, 1946-1948, p. 293; Kondrašov, 1974, p. 179).

On enlarging the meaning, the principle of transfered naming of phenomena was applied. The most used level of language was spoken one, or more precisely folk lexis, e.g. *kluť sa* ("appear, turn up"), *roztrhať sa* ("to fall apart, to cease to exist by breaking internal bonds", e.g. alliance), *rozpadať sa* ("to split up, to sectionalize into the several parts"), *klesnúť* ("to retreat, to become extinct"), *vitĭect'* ("to unfold") and others. The same principle was often applied within nouns, e.g. *daromňica* ("trifle"), *štĭep* ("part, section"), *štrbina* ("blank" abstract meaning), *haluz hospodárstva* ("sector"), *čuridlo* ("caricature"), *zásťera* ("cloak") and likewise.

Noticeable prevalence of productivity of abstracts is typical mainly of forming new namings in the area of the so-called cultural lexis, namely expressions generated by suffix *-osť* and expressions generated by the sufix *-ja/-je.* The frequent use of word formation of the type *-osť* is evidenced also by using a pronoun as a motivating element if it had an adjective form, e.g. *kolkosť* (quantity), *všetkosť.* When it comes to verbs, it was mainly expressions generated by prefixes. Nearly all the prefixes that still exist in the contemporary Slovak are held efficient. In Štúr's Slovak e.g. such prefixed verbs were formed: *dotvrdzuvať* (something), *namožďiť sa* ("to worry"), *nadužívať, odmieňať sa* ("to take turns"), *prečuduvať sa, rozdumať sa, sprofanuvať sa, vinachádzať sa* ("to occur"), *zčlánkovať sa* ("to unite ") and likewise.

In terms of suffixes, the most frequently used was *-uvať.* For instance, *premestuvať sa* (somewhere), *dozveduvať sa*, *podsekuvať* (something); highly productive was mailny by forming verbs out of nouns: *amalgamizuvať, anatomizuvať, assimiluvať, kandiduvať and others.*

As far as transference of foreign words is concerned, there are borrowed words mostly from Latin, Greek, less from German and Slavic languages – Czech, Russian, Polish, Serbo-Croatian language in Štúr's Slovak (in detail Kondrašov, 1974, p. 209 – 242). In terms of genesis of respective words, we can speak of a specific relationship between the Czech and Slovak language. According to A. N. Kondrašov (ibid. P. 210) it is easy to determine the origin of Czech words only within the so-called cultural words that originated in the first half of the 19th century in Czech and they were then downloaded to Slovak. In early Štúr's lexis, there are numerous Czech lexical items, or more precisely names related to Czech words (e.g. *potkať, sklíčiť niekoho* "to deject" , *výmluvnosť, dúkaz, rostlina, posvecení* and so on.) There is a number of loan words from Czech in the field of technical terms.

Translation of foreign words was often used to enlarge the number of neologisms while the original words were written in brackets in order to explain the meaning, e.g.: *Middle Ages (Miitelalter), jet (Einfluss), bill (Wechsel), minority (Minorität), telegraphy (Telegraph), derivation (Derivatio), fraction (Fractia) and so on.*

Typical feature of Štúr's lexis was considerable polylexis. Its accompaniment was a bigger competition among synonymic words, to a great extent among word forming synonyms. It was usually about the competition among word forming types within particular word formation categories. For example, this way there were formed synonymous pairs: *posilňeňje – posila, vihováraňja – víhovorka, boleňje – boľasť, zasednutje – sedňica;* by naming the result of the plot: *staveňje/stavaňje – stavba* (in more detail in: J. Furdík, 1968), then word forming synonyms were created for example by naming things: *majetok – majetnosť,* of things*: bažnosť – bažívosť*, factors of the plot: *spevec – spevák* and so on.

As a result of the enlarged lexical variety, new word forming synonymous lines are being created, and that either as absolute synonyms, e.g. *prípad – prípadok – prípadnosť*, or with not absolutely identical semantic features, e.g. *cestuvateľ* („cestujúci") – *cestovník* (e.g. on board ship) – *cestujúci* ("traveller"). A specific case are verbs with different preposition phrases, but with the same meaning, e.g. *pustiť sa* k niečomu/na niečo, *bažiť* po niečom/za niečím, *ujednostajňiť sa* o niečom/na niečom, *zažalostiť* nad niečím/za niečím, *opierať sa* na niečo/o niečo, *natrafiť* niečo/na niečo. Zaznamenali sme aj slovesá s tromi variantnými pádovými väzbami, napr. *diviť sa* nad niečím/na niečom/niečomu, *oddať sa* na niečo/do niečoho/niečomu, *zaujímať sa* o niečo/za niečo/niečoho.

Under the given contexts, we can state that lexis of Štúr's standard Slovak was characterized by distinct rise of neologisms that appeared throughout the whole journalism of that period. Their number is undoubtedly a sign of a unique extent of modernization of vocabulary within an uncommonly short time and moreover on a relatively small text space.

2 The dictionary project

The listed brief characteristics of the lexis of Štúr's Slovak indicates that its users had a very inventive, direct approach to the language as a functional and sufficiently shapable means of communication. Štúr's neologisms represent a peculiar chapter in the history of standard Slovak, which undoubtedly deserves its own lexicographic processing. Yet if we today look back at the whole period of functioning of Slovak as a standard, let us say, literary language, we can see that the period of Štúr's standard Slovak is least lexicographically processed (a bilingual dictionary by Š. Jančovič: *Noví slovensko-maďarskí a maďarsko-slovenskí slovňík* was published in two editions in 1848). The dictionary of Štúr's Slovak (DSS) is a long planned project at Nitra's Department of Slovak language (comp. Ľ. Kralčák: The project of Štúr's Slovak and its computer support, 2001). At the moment this project is becoming more viable, mainly after becoming a guaranteed VEGA project in 2008.

Already on designing the first draft of the project DSS (comp. Kralčák, ibid.) we expected three areas of work: collection of language material, lexicographical analysis of the texts and creation of the lexical database. In the current project there is an

ambition to set up at least a part of the core DSS in the work form. Project work is divided into two stages.

2.1 The main objective of the first stage is to collect language material thusly making a base for the incoming lexicographical work. This effort has two sides: linguistic and informatic.

2.1.1 The linguistic side od collecting the language material includes selection of the source base of the dictionary, that is inventory of the source texts. The main aim of DSS is to record the vocabulary of Štúr's period as a treasured cultural heritage to the maximum extent. Therefore it is necessary to solve the question of the source base of the dictionary according to the "maximalistic model" and that in its vertical delimitation (Štúr's texts from the broadest period - it is a matter of periodization of Štúr's Slovak; in our project this period consists of texts from 1844 to 1852), and in horizontal delimitation (all types of texts from the listed period, hence texts representing all the developing functional styles). Yet the maximalistic model does not mean that we will process the complete inventory of the texts of Štúr's Slovak. After considering the possibilities of the project team, we made a decision not to include here par excellence texts that were only manuscripts. This does not also include private correspondence of Štúr's colleagues but neither texts such as *Slovo za slovenčinu* by C. Zoch, drama (fragment of the text) *Odchod z Bretislavi* by M. Dohnány and others. The source base is also made by periodicals (*Slovenskje pohladi na vedi, umeňja a literatúru, Slovenskje národňje novini,* their literary insertion *Orol Tatránski*, almanach *Ňitra*), Štúr's authorial print, then Francisci, Hodža, Kadavý, Hurban, Dohnány, Sládkovič (authorial texts are naturally numerous in the above mentioned periodicals). The listed texts simultaneously represent the most important production from all the developing functional styles in Štúr's period.

2.1.2 Informatic side assumes the creation of the lexical database with the help of the electronic record of the texts. It is the initial step towards the creation of the text corpus of Štúr's Slovak. The preparatory project already mentions that there are two possibilities how to transcribe the texts into the electronic form, namely by the use of a text editor or a visual scanner of letters. We have opted for the simple transcription as the visual scanning loses its efficiency due to numerous mistakes which occured mainly as a result of insufficient quality of the print of the processed material.

2.2 The main objectives for the second stage are as follows:

- lexicographic analysis of the obtained material,
- formation of the corpus of the dictionary of Štúr's Slovak in the working form.

At this stage we will use the electronic corpus of the texts and by the use of certified software (work with the program WordSmith) we will proceed to the creation of the lexical database and its follow-on lexicographic processing. With respect to the fact the work carried out on the project requires a longer lasting research and collecting the language material by more co-workers. The formation alone will be executed as an output within the stated project only in the working form, or better said depending upon the financial support also in the form of the so-called sample exercise book.

Other works after completing this stage of the project concentrate on the gradual publishing of the dictionary.

The conclusion. – At this point the project *Dictionary of Štúr's Slovak* is being executed by the team of co-workers at Nitra's Department of Slovak language. The project team is presently working on the first stage of the project, that is we focus on collecting the language data in the electronic form so that we can proceed to its lexicographic processing in the following stage.

References

FURDÍK, Juraj: O tvorení názvov deja a výsledku deja v štúrovskom období. In: Jazykovedný zborník venovaný VI. slavistickému kongresu. Acta Facultatis Philosophicae Universitatis Šafarikanae Prešovensis. Bratislava, SPN 1968, s. 89–106.

HORECKÝ, Ján: K charakteristike štúrovského lexika. Linguistica Slovaca, 4 – 6, 1946 – 1948, s. 279 – 298.

KONDRAŠOV, Nikolaj Andrejevič: Vznik a začiatky spisovnej slovenčiny. Bratislava, Veda 1974. 284 s.

KRALČÁK, Ľubomír: Projekt slovníka štúrovskej slovenčiny a jeho počítačová podpora. Bratislava: VEDA 2001, s. 150 – 154.

Sources for the Dictionary of Štúr's Slovak Project

Journals:

Slovenskje pohladi na vedi, umeňja a literatúru. 1846.
Slovenskje pohladi na vedi, umeňja a literatúru. 1847.
Slovenskje pohladi na vedi, umeňja a literatúru. Ďjel I., svazok 3. 1851.
Slovenskje pohladi na vedi, umeňja a literatúru. Ďjel II.Svazok 1- 6. 1851.
Slovenskje pohladi na vedi, umeňja a literatúru Ďjel III. 1852.
Slovenskje pohladi na vedi, umeňja a literatúru Ďjel IV. 1852.
*Slovenskje národňje novini.*1845, č. 1 - 44.
Slovenskje národňje novini. 1846, č. 45 - 147.
Slovenskje národňje novini. 1847, č. 148 - 248.
Orol Tatránski. Ročník I. 1845/1846.
Orol Tatránski. Ročník II. 1846/1847.
Orol Tatránski. 1848. Ročník III. 1847/1848.
Ňitra. Dar drahím krajanom slovenskím obetuvaní. Ročník II. 1844.
Ňitra. Dar drahím krajanom slovenskím obetuvaní. Ročník III. 1846.
Ňitra. Dar drahím krajanom slovenskím obetuvaní. Ročník IV. 1847.
Novini pre hospodárstvo, remeslo a domáci život. 1848.

Other sources:

LICHARD, D.: *Domová pokladnica na rok običajní 1847.* Skalica. Ročňík prví.
Domová pokladnica na rok prjestupní 1848. Skalica. Ročník druhí.
Domová pokladnica na rok običajní 1849. Skalica. Ročník treťí.
DOHNÁNY, M.: *Historia povstaňja z roku 1848*. Skalica 1850.
ŠTÚR, Ľ.: *Nárečja slovenskuo alebo potreba písaňja v tomto nárečí.* V Prešporku 1846. *Nauka reči slovenskej*. V Prešporku 1846.
FRANCISCI, J.: *Slovenskje povesťi*. V Levoči 1845.
HODŽA, M. M.: *Dobruo slovo Slovákom súcim na slovo*. Levoča 1847.
SLÁDKOVIČ, A.: *Marína*. Bratislava 1846.
HODŽA, M. M.: *Ňepi pálenku to je ňezabi.* B. Bystrica 1845.
DOHNÁNY, M.: *Podmaňínovci.* Činohra v 4. ďejstvách. Levoča 1848.
HURBAN, J. M.: *Slovo o spolkách mjernosťi a šklách ňeďelních.* B. Bystrica 1846. *Českje hlasi proťi slovenčiňe.* Skalica 1846.
KADAVÝ, J.: *Čítanka pre malje ďjetki.* Buďín 1845. *Prjaťel ludu. Knižka pre slovenských hospodárou a remeselňíkou*. Buďín 1845.
JANČOVIČ, Š.: *Noví slovensko-maďarskí sa maďarsko-slovenskí slovňík. I. oddďjel. Slovensko-Maďarskí slovňík. II. oddďjel*. Sarvaš 1848.

Annotation Procedure in Building the Prague Czech-English Dependency Treebank

Marie Mikulová and Jan Štěpánek

Institute of Formal and Applied Linguistics
Charles University in Prague, Czech Republic

Abstract. In this paper, we present some organizational aspects of building of a large corpus with rich linguistic annotation, while Prague Czech-English Dependency Treebank (PCEDT) serves as an example. We stress the necessity to divide the annotation process into several well planed phases. We present a system of automatic checking of the correctness of the annotation and describe several ways to measure and evaluate the annotation and annotators (inter-annotator accord, error rate and performance).

1 Introduction

Building a huge corpus with rich linguistic annotation calls for elaborate organization of the annotation process. In our contribution, we will present such a project, namely Prague Czech-English Dependency Treebank (PCEDT). In the first place, we will focus on the organizational aspects of the building of the corpus that can be generally applied to building of any similar huge corpus. In particular, the main points will be:

- division of the annotation into several phases
- system for checking the accuracy of the annotation
- ways of evaluation of the annotation and annotators

PCEDT is planned to be a corpus of (deeply) syntactically annotated parallel texts (in English and Czech) intended chiefly for machine translation experiments. The texts for PCEDT were taken from Penn Treebank [4], which means there are mostly economical articles from the Wall Street Journal. 2312 documents were used in PCEDT (approximately 49,000 sentences) that are manually annotated with constituent trees in Penn Treebank. For the Czech part of PCEDT, the English texts were translated into Czech.

As a base of the process of creation of the corpus (hierarchical system of annotation layers, annotation rules) we will use the already accomplished Prague Dependency Treebank (PDT) 2.0 [1]. While organizing the annotation of PCEDT (especially its Czech part, which is the main concern of this article), we will prop ourselves upon multifarious experiences (both positive and negative) gained from the production of PDT 2.0.

2 Division of the annotation into several phases

If one builds up a corpus in which a rich and complex linguistic information is attached to the input data (i.e. sentences), according to our experience it is advisable to divide this process into several partial phases. The question how to divide the annotation when the information attached is mostly very complex and various phenomena are interconnected remains rather difficult.

Example of such a rich annotation is the tectogrammatical (deep syntactical) layer of PCEDT (and similarly the same layer of PDT 2.0). In the annotation process, each sentence is assigned a deep syntactical structure (which among others deals with ellipsis and valency of verbs and nouns); each unit of the structure is assigned its deep syntactical function (there are several tens of the "functors") and many attributes, mainly grammatemes (tectogrammatical counterparts of morphological categories). The tectogrammatical tree captures coreference, topic-focus articulation and deep word order. There are 39 different attributes, for a node of a tectogrammatical tree in PDT 2.0 there are 8.42 attributes filled on average.

The annotation of the tectogrammatical layer was divided into several phases when the layer was being created for the PDT 2.0 already. The division is inevitable because no annotator is able to keep all the annotation rules for all the annotated phenomena in his or her head (the annotation manual [3] has more than 1000 pages). Moreover, the more information the annotator has to attach to the data, the more likely he or she omits some of the details.

The experience from production of PDT 2.0 unveiled that the division of the annotation process into several steps is desirable for the quality of the output data, even if some phenomena had to be reconsidered repeatedly by different annotators in various phases.

Today, when building PCEDT, we increased the number of phases even more. The first phase of the PDT 2.0 annotation (creating a syntactical structure and assigning functions to the units of the structure) proved to be still too complex and comprising too many features. We also tried to get rid of repeated resolution of the same problems, e.g. by introducing new temporary values for some attributes whose final values will be judged in a later phase of the annotation.

For the tectogrammatical annotation, we count on the following phases:

1. building a tree structure, dealing with ellipsis included; assignment of functors and valency frames, links to lower layers (10 attributes).
2. annotation of subfunctors (fine grained classification of functors, 1 attribute).
3. annotation of coreference (4 attributes).
4. annotation of topic-focus articulation, rhematizers and deep word order (3 attributes).
5. annotation of grammatemes, final form of tectogrammatical lemmata (17 attributes).
6. annotation of remaining phenomena (quotation, named entities etc.)

The first phase is still the most difficult, each annotator is responsible for the whole structure of the tree and correct values of ten attributes. All these attributes are connected with the structure and deep syntactical functions of the nodes. The annotator does not have to pay attention to anything else.

For the first phase of the annotation process a "working value" *#NewNode* was established for tectogrammatical lemmata of nodes added to trees in case of ellipsis of valency frame arguments. Absent obligatory arguments are represented by added nodes in final tectogrammatical trees and their tectogrammatical lemmata signify the type of the elision (*#Gen* stands for a general participant, *#PersPron* for a deletion, *#Cor* and *#QCor* for a controlee in control constructions, *#Rcp* for ellipses because of reciprocation). The type of elision is closely connected with coreference (some types of absent arguments have a coreferent, some do not). During the annotation of PDT 2.0, the lemmata of absent arguments were assigned in the first phase, the coreference in a following phase. By introducing the *#NewNode* value the final solution of the tectogrammatical lemma was postponed to the following phase together with remaining questions of coreference. An annotator inserts a node with the "working value" of the tectogrammatical lemma and only assigns its syntactical function, not taking care about the lemma.

In the first phase we also "neglect" the annotation of rhematizers (in PDT 2.0, they were annotated in the first phase). Competent decision about a rhematizer (whether an expression is a rhematizer or not, its position in the tree) is possible only if the topic-focus articulation of the sentence is decided at the same time. Therefore, definitive annotation of rhematizers is planned to the topic-focus annotation phase.

Determine an amount of annotated information that does not harm the quality of the data is obviously difficult. Our believe that the current schedule of phases constitutes a reasonable and manageable rate seems to be justified by the measuring of inter-annotator agreement (see section 4.1). The quality of the data is regularly guarded by a system of automatic checking procedures (see section 3).

3 System for checking the accuracy of the annotation

When PDT 2.0 was in production, only random "manual" checks of the accuracy were performed. The real checking took place when all the annotation had finished. The checking and fixing phase was quite complex and time-consuming; moreover, in some cases, the changes were not realized full-scale [5].

We want to avoid such a procedure in the development of PCEDT. Checking of the data is performed in parallel with the annotation process. At the beginning of the process, a number of automatic checking procedures was proposed and new tests subsequently come up during the annotation process. Currently there are 99 checking procedures that verify the annotation of Czech sentences.

The checking procedure proposal is based on the fact that many annotation rules imply that particular phenomena cannot (or have to) occur in the annotation output. They mainly combine attribute values and structure of a tree. For example, a simple

check states that every coordination has at least two members and reports all one-member coordinations as errors. Another check states that the root of a tectogrammatical tree has only a limited set of possible functors (PRED for a predicate, DENOM for nominative clause, PARTL for interjection clause etc.). There is also a converse check monitoring that no dependent node has the PRED functor, and so on.

The checking procedures return a list of erroneous (questionable) positions in the data. The annotator gets his or her data back for corrections, manually fixing each position.

The checking procedures are run periodically after a given volume of the data has been annotated (1000 sentences) or once a quarter. All the data are checked every time (in case a new check existed) and after the correction, the data are checked again and again while there are any errors (new errors can arise in fixing the old ones).

Automatic checking procedures improve the quality of the data not only by fixing the present errors, but also by providing a feedback to the annotators (because each annotator fixes his or her own data, i.e. his or her own errors) and thus eventually improving the future annotation.

4 Ways of evaluation of the annotators

A system for evaluation of the annotation and annotators should by an integral part of any annotation project. In the PCEDT project, the quality of the work of a particular annotator is judged by several ways:

- the annotation agreement between annotators is measured,
- the output of the automatic checking procedures tells us how often an annotator makes mistakes compared to the others,
- the annotators book the time they spend annotating; it allows later to evaluate their performance and the relation of the efficiency to the error rate.

4.1 The annotation agreement between annotators

The basic way how to evaluate an annotation is to measure the inter-annotator agreement. However, the structure to be compared is very complex. The algorithm aligning two tectogrammatical trees built upon the same analytical tree is described in detail in [3], once the trees are aligned node to node, we just compare the values of all the attributes of all the aligned nodes. To evaluate the structural agreement, we treat the identifier of a node's parent as a new attribute of the node. Complex attributes (lists, structures etc.) need further manipulation in order to be compared. For example, identifiers of linked analytical nodes have to be sorted; for annotator's comment, we only compare the type, because the text can vary.

Since there is no “golden” annotation, we just measure the agreement of all the pairs of annotators (see Table 1, data from December 2007; average value is shown for every attribute, and average value over all the attributes and structure is presented as “Overall”). As a baseline, we use the output of an automatic procedure with which the annotators start their work (marked “Z” in the table). Note that the

agreement among annotators is always higher than the agreement between any annotator and the baseline. The attributes with a lower difference between baseline and the annotators (about 5%, i.e. *is_state*, *is_generated*, *is_dsp_root*, *compl.rf*, *annot_comment,* and *a/lex.rf*) tend to contain more errors, or have too vague annotation rules.

The annotator that agrees most with all the others ("K") is at the same time the annotator that makes the least errors and submits the most sentences (see next sections).

Overall	K	94,08%			
	Ma	94,01%			
	A	93,83%			
	O	93,78%			
	Z	84,58%			
Structure	A	88,62%	is_dsp_root	K	95,86%
	Ma	88,60%		A	95,83%
	O	87,92%		Ma	95,75%
	K	87,88%		O	95,72%
	Z	69,28%		Z	89,72%
a/aux.rf	K	93,82%	is_generated	K	96,24%
	Ma	93,58%		A	96,05%
	A	93,55%		Ma	96,03%
	O	93,53%		O	96,02%
	Z	82,45%		Z	90,27%
a/lex.rf	K	96,26%	is_member	K	94,72%
	Ma	96,12%		A	94,70%
	A	96,00%		Ma	94,50%
	O	95,90%		O	94,25%
	Z	89,67%		Z	85,47%
annot_comment	K	96,52%	is_parenthesis	Ma	95,42%
	Ma	96,40%		K	95,40%
	A	96,30%		O	95,27%
	O	96,27%		A	95,15%
	Z	90,43%		Z	88,72%
compl.rf	K	96,32%	is_state	K	96,50%
	Ma	96,22%		Ma	96,25%
	A	96,12%		O	96,13%
	O	96,03%		A	96,13%
	Z	90,18%		Z	90,35%
functor	K	85,70%	t_lemma	K	93,76%
	Ma	85,67%		Ma	93,60%
	O	85,57%		O	92,70%
	A	85,13%		A	92,42%
	Z	66,80%		Z	81,60%

Table 1. Inter-annotator agreement

4.2 Error rate

Using the list of errors generated by the checking procedures we can count how often the annotators make errors (only those errors the procedures can detect, of course): the number of errors the annotator made is divided by the number of sentences or nodes she annotated. Table 2 shows the comparison of the error rate for 4 annotators in December 2007 (at the beginning of the process) and current numbers for 7 annotators from July 2009. The numbers from different periods cannot be compared directly because since the beginning there have been more than 30 new checking procedures, which means the current list of errors is longer. On the other hand, the rank of the annotators can be compared.

The table shows that our current best annotator ("K") had approximately 30 errors per 100 sentences and 1.62 errors per 100 nodes. Her error rate has not got worse over the two years and she remains the best annotator. The table further shows that the differences in error rate between annotators can be great and that all the annotators keep their positions: no one gets markedly better nor worse. The comparison of veteran annotators and the new ones that annotate only for a short time is also interesting: it shows that knack, practice, and experience lead to quality of the annotation.

	December 2007		July 2009	
Who	Errors per 100 sentences	Errors per 100 nodes	Errors per 100 sentences	Errors per 100 nodes
K	29.7851	1.6241	1.5103	0.0806
O	39.6699	2.0624	4.0331	0.2067
Ma	61.4087	3.2707	8.4670	0.4533
A	63.2318	3.3498	6.3583	0.3265
L	-	-	15.0668	0.8010
Mi	-	-	16.2241	0.8460
J	-	-	19.0476	1.0971

Table 2. Error rate

4.3 Performance of the annotators

In the annotation process, even the time spent working by the annotators is measured. The annotators book the time to a web form. For each month the web application counts the annotators' performance over the month and the over-all performance. The data are important among others to determine the wages; on the basis of the data we tariff a sentence (annotators are being paid monthly according to the number of sentences they have annotated).

Table 3 shows performance of the annotators in October 2008 and June 2009, table 4 shows the over-all performance. Monitoring the performance illustrates the differences between annotators, but also the fluctuation of each particular annotator. We can also observe the inverse proportionality of the performance and error rate (see section 4.2): the more is the annotator efficient (she annotates more data), the less errors she makes.

	October 2008				June 2009			
Who	Hours	Sentences	Sentences per hour	Minutes per sentence	Hours	Sentences	Sentences per hour	Minutes per sentence
A	18.50	147	7.946	7.551	-	-	-	-
I	100.50	742	7.383	8.127	101.50	1229	12.108	4.955
J	11.50	97	8.435	7.113	2.00	28	14.000	4.286
K	33.00	418	12.667	4.737	23.50	332	14.128	4.247
L	46.00	143	3.109	19.301	27.88	365	13.092	4.583
Ma	40.00	310	7.750	7.742	-	-	-	-
Mi	17.85	142	7.955	7.542	24.91	358	14.372	4.175
O	37.81	403	10.659	5.629	56.65	632	11.156	5.378

Table 3. Performance of the annotators

Who	Hours	Sentences	Sentences per hour	Minutes per sentence
A	114.25	963	8.4289	7.1184
I	827.00	7006	8.4716	7.0825
J	105.70	1001	9.4702	6.3357
K	107.00	1430	13.3645	4.4895
L	266.41	1716	6.4412	9.3150
Ma	78.00	615	7.8846	7.6098
Mi	169.98	1655	9.7364	6.1624
O	289.02	3211	11.1100	5.4006

Table 4. Over-all performance of the annotators

5 Conclusion

In the article, we have presented some organizational aspects of building of a large syntactical treebank. We stressed mainly the necessity to divide the annotation process into several well planned phases. We presented out system for checking the correctness of the annotation. The fact that the correctness is being checked at all should be pointed out: it is not a common practice in similar projects. We described three ways to measure and evaluate the annotation and annotators.

We believe that having published PDT 2.0 with 50,000 sentences annotated on the tectogrammatical layer and being in the halftime of the PCEDT project with more than a half data already annotated (33,500 sentences, 68% of the corpus) our proposals are sufficiently backed by our experience and practice.

Acknowledgement:

The research reported in this paper was supported by the LC536, GAUK 22908/2008, and FP6-IST-5-034291-STP.

References

[1] Hajič, J. et al. (2006). The Prague DependencyTreebank 2.0. CD-ROM. Linguistics Data Consortium Cat. No. LDC2006T01. Philadelphia, PA, USA. URL: `http://ldc.upenn.edu`, `http://ufal.mff.cuni.cz/pdt2.0`

[2] Klimeš, V. (2006). Analytical and Tectogrammatical Analysis of a Natural Language. PhD Thesis. MFF UK, Prague.

[3] Mikulová, M. et al. (2006). Annotation on the tectogrammatical level in Prague Dependency Treebank. Annotation manual. Technical report ÚFAL TR-2006-30. MFF UK, Prague.

[4] Mitchell, M. et al. (1995). Treebank 2. CD-ROM.. Linguistics Data Consortium Cat. No. LDC95T7. Philadelphia, PA, USA. URL:
`http://ldc.upenn.edu`
`http://www.cis.upenn.edu/~treebank/`

[5] Štěpánek, J. (2006). Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konsistence dat) [Capturing a Sentence Structure by a Dependency Relation in an Annotated Syntactical Corpus (Tools Guaranteeing Data Consistence)]. PhD Thesis. MFF UK, Prague.

Automatic Analysis of Terminology in the Russian Corpus on Corpus Linguistics

Olga Mitrofanova[1] and Victor Zakharov[1,2]

[1] Saint-Petersburg State University
[2] Institute of Linguistic Studies of the Russian Academy of Sciences, Russia
alkonost-om@yandex.ru vz1311@yandex.ru

Abstract. The paper is devoted to processing terminological items occurring in domain-restricted texts. Linguistic resource involved in research is the Russian corpus on corpus linguistics developed in St. Petersburg State University and Institute of Linguistic Studies, RAS. (Semi-)automatic terminology extraction is performed with the help of linguistic and statistical tools which allow to generate lists of single-word and multi-word terms supplied with frequency data and lexical-syntactic patterns. Lexical-syntactic patterns are used in the analysis of contexts which contain definitions of terms, expose interrelations between terms, provide their synonyms, translation equivalents, etc. Results obtained contribute much to the construction of the Russian thesaurus on corpus linguistics.

1 Introduction

Domain-restricted text corpora reflecting knowledge on particular subject areas stand out against a background of the variety of corpus genres. The given type of corpora is distinguished by rigid restrictions in kinds and topics of texts, formalization of text content due to logical and conceptual structure of subject areas; transparent hierarchical structure of the lexicon owing to its saturation with terms; the influence of scientific content and style on lexical, semantic, morphological, syntactic parameters of corpus texts [Gerd 2005].

Results of the analysis of corpora developed for particular subject areas are of great practical importance. Domain-restricted text corpora and linguistic data extracted from them are used both in scientific and technical lexicography (in compiling terminological dictionaries, classifiers, etc.) and in the field of automatic text processing (automatic indexing, summarization and classification, information retrieval, machine translation, etc.). Domain-restricted text corpora serve as a basis for (semi-)automatically created terminological dictionaries and thesauri, formal ontologies, mono- and multilingual resources.

Research in the field of development and further application of domain-restricted text corpora created for new subject areas attract much attention of linguists and terminologists. Certainly, corpus linguistics itself belongs to such new areas.

Complex description of corpus linguistics terminology has been performed for English [Baker et al. 2006] as well as for some other languages including Slavic, cf.

corpus linguistics section of the Slovak terminological database which was built in Ľ. Štúr Institute of Linguistics, the Slovak Academy of Sciences, Bratislava [Levická 2007; Šimková 2006] (URL: `http://data.juls.savba.sk/std/`). However, until recently, this subject area was not presented in Russian terminology.

The main tasks of the project carried out by the Department of Mathematical Linguistics, St. Petersburg State University and Institute of Linguistic Studies of the Russian Academy of Sciences are the creation of the Russian corpus on corpus linguistics, the Russian thesaurus on corpus linguistics and development of a range of linguistic resources on its basis.

The present paper discusses particular aspects of the project, namely, automatic extraction, processing and systematization of terms from the Russian corpus on corpus linguistics.

2 The Russian corpus on corpus linguistics and conjoined linguistic resources

The Russian corpus on corpus linguistics includes numerous text reflecting a wide range of corpus linguistics problems, among them definition of corpus linguistics as a particular sphere of research activity; its opposition to other divisions of linguistics and language engineering; definition of a corpus with respect to other kinds of linguistic data; procedures of creating and using corpora (tagging, alignment, searching, etc.); corpus typology; developers and users of corpora; interrelations between corpora and other linguistic resources.

The core part of the Russian corpus on corpus linguistics consists of research papers in Russian: conference proceedings [Corpora 2002, 2004, 2006, 2008, etc.], separate articles, manuals, textbooks, monographs, etc. The corpus is regularly updated and enriched with new materials.

Each text of the corpus is preprocessed: special tags for tables, images, formulae, links, numbers, non-Russian text fragments, etc. are introduced, morphological tagging and lemmatization are performed. Further the texts are supplied with metadata (metatagging) which include bibliographic passport and a set of terms (key words) indicating the topic of the paper.

Cluster analysis of terms (key words) [Mitrofanova, Panicheva, Savitsky 2007] allowed to form the nuclear part of formal ontology which reflects major concepts of corpus linguistics [Vinogradova, Mitrofanova 2008]. Expansion of formal ontology implies automatic retrieval and further analysis of terms occurring in the processed texts. Among other tasks of the project are thematic classification of texts, construction of the Russian thesaurus on corpus linguistics, etc. To fulfill these tasks it is necessary to perform (semi-)automatic extraction of terms from the corpus.

3 Techniques used for term extraction and Alex+ terminological toolkit

Major approaches to term extraction often require linguistic and/or statistical analysis of corpora. Linguistic techniques of term extraction imply mainly manual text processing performed by experts who select supposed single-word and multi-word terms. Statistical techniques treat terms as frequent lexemes or word-groups. Multi-word terms are regarded as n-grams characterized by high values of association measures (e.g. *MI-score*, *t-score*, *Log-Likelihood*, *C-value*, χ^2, etc.).

Combination of linguistic and statistical techniques seems to be highly productive in the analysis of terminology: cf. research results dealing with Slavic languages, Russian in particular [Braslavskij, Sokolov 2008; Dobrov et al. 2003; Kupść 2007; Urbańska, Piechociński 2007]. Hybrid approach to term extraction deals with (semi-)automatic processing of domain-restricted corpora and require the use of lexical-syntactic patterns, various filters (e.g. stop-word dictionaries) alongside with frequency analysis and collocation extraction [Zakharov, Khokhlova 2008].

Various toolkits aimed at (semi-)automatic term extraction have been developed, among them Alex+ for Russian [Sidorova 2008]. Alex+ turns out to be a convenient lexicographic environment for development and maintenance of domain-restricted dictionaries and thesauri which allows to process corpora, extract single-word and multi-word terms from texts taking into account certain lexical-syntactic patterns, perform statistical analysis of texts in corpora, automatically enrich dictionaries using learning samples. Alex+ includes modules for morphological analysis, assemblage of multi-word terms according to particular lexical-syntactic patterns, concordance creation and revision, stop-words detection, etc. Dictionaries developed with the help of Alex+ include various types of information about terminology (domain data, semantic features, concept hierarchy, statistical features, etc.). Alex+ allows to develop formal ontologies alongside with dictionaries and thesauri which can be used simultaneously for classification of texts in corpora. Thus, possibilities of Alex+ fully correspond to the tasks of research project in question, that's why it was used for (semi-)automatic extraction of terms from the Russian corpus on corpus linguistics.

4 Research results

4.1 Description of single-word and multi-word terms with the help of lexical-syntactic patterns

Texts included in the Russian corpus on corpus linguistics were processed with the help of Alex+ so that lists of single-word and multi-word terms were successfully formed and sorted with regard to general lexical-syntactic patterns, frequency, etc.

It is very important to distinguish carefully terms and non-terms. On the one hand, one should rely upon expert opinions; on the other hand, there are certain criteria allowing to solve the problem, e.g. specificity criterion [Šajkevič 2003].

Lists of terms were edited: stop-words (function words) were marked out and added to stop-word dictionary, non-terms (e.g. *miro*) were deleted.

List of single-word terms include nouns, adjectives and verbs, e.g.:
N: *vydača, dannyje, dokument, zapros, lemma, metka, razmetka, častota,* etc.;
Adj: *avtomatizirovannyj, informacionno-poiskovyj, korpusnoj, korpusnyj,* , etc.;
V: *avtomatizirovat', razmečat'*, etc.

List of multi-word terms include word-groups representing various lexical-syntactic patterns, e.g.:
Adj+N: *avtomatičeskaja obrabotka / razmetka / systema, avtomatičeskij analiz / režym, anaforičeskaja / morfologičeskaja / semantičeskaja / sintaksičeskaja / strukturnaja /prosodičeskaja razmetka,* etc.;
N+N: *korpus dannych / tekstov, model' jazyka, razmetka korpusa / dokumenta / teksta, razmer korpusa, raspoznavanije reči,* etc.;
N+Prep+N: *dostup k korpusu, nauka o jazyke, poisk v korpuse,* etc.;
N+Adj+N: *bank sintaksičeskich struktur, massiv jazykovych dannych, obrabotka tipovych zaprosov,* etc.;
N+Prep+Adj+N: *korpus s sintaksičeskoj razmetkoj, teksty na jestestvennom jazyke,* etc.;
Adj+N+N: *avtomatičeskaja obrabotka teksta, kompjuternaja baza dannych, kompjuternaja model' jazyka,* etc.;
N+N+N: *vyvod rezul'tatov poiska, standart predstavlenija metadannych,* etc.;
N+Prep+N+N: *poisk s ukazanijem konteksta,* etc.;
Adj+Adj+N: *ustnaja razgovornaja reč ,* etc.

The most frequent lexical-syntactic patterns are **Adj+N**, **N+N**, **N+Adj+N**, **Adj+N+N**, **N+N+N**.

Multi-word terms exposed and described with the help of Alex+ differ not only in structure and content but also in cohesion, especially tree- and four-word terms, as they are supposed to be combined of one- and two-word terms. Level of cohesion can be evaluated in course of statistical analysis.

4.2 Description of terminological contexts with the help of lexical-syntactic patterns

Lexical-syntactic patterns are successfully used in the analysis of terminological contexts which contain definitions of terms, expose interrelations between terms, provide their synonyms, translation equivalents, etc. Annotation of terminological contexts allows to point out special context markers which can be used in selection of terms from the processed texts and revelation of their significant features. The structure and typical content of terminological contexts may be described by various lexical-syntactic patterns, e.g.:

NP(def) <*nazyvat' / nazyvat's'a / imet' nazvanije*> NP(term)
(definitions of terms):
Eto kodirovanije informacii ***imejet nazvanije*** *metarazmetka*… [Zakharov 2005];
NP(term) <, *ili*> NP(term)
(synonymic relations between terms):
…*sintaksičeskogo analiza*, ***ili*** *parsinga*…[Zakharov 2005];
NP(term) <*javl'at's'a rezul'tatom*> NP(term)
(«process – result» relations between terms):
……*sintaksičeskaja razmetka*, ***javl'ajuš'ajas'a rezul'tatom*** *sintaksičeskogo analiza, ili parsinga*… [Zakharov 2005];
NP(term) <*obespečivat'*> NP(term)
(«item – purpose» relations between terms):
…konvertirovanije razmečennych tekstov v strukturu specializirovannoj lingvističeskoj informacionno-poiskovoj sistemy (corpus manager), ***obespečivajuš'ej*** *bystryj mnogoaspektnyj poisk i statističeskuju obrabotku* … [Zakharov 2005].

Analysis of terminological contexts contributes much to exposure of term hierarchy, specification of the word list for the developed thesaurus, enrichment of the thesaurus with definitions. Definition module of the thesaurus includes definitions extracted directly from texts of the corpus and definitions constructed according to regular lexical-syntactic patterns occurring in the corpus. The use of lexical-syntactic patterns provides uniformity of descriptions given to terms in the thesaurus. In future it seems reasonable to use special language for annotation of lexical-syntactic patterns, e.g. LSPL (Lexical-Syntactic Pattern language) [Bolšakova et al. 2007; Rabčevskij et al. 2008].

5 Conclusion

The given research allowed working out the strategy of (semi-)automatic terminology extraction from the Russian corpus on corpus linguistics. Alex+ terminological toolkit was used for single-word and multi-word term extraction. Major lexical-syntactic patterns of single-word and multi-word terms were revealed and described. Terminological contexts occurring in the corpus were analyzed with the help of enriched lexical-syntactic patterns as well.

Results of automatic terminology extraction contribute much to the construction of the Russian thesaurus on corpus linguistics. It is assumed that the developed terminological thesaurus should inherit the hierarchy of concepts reflected in the formal ontology on corpus linguistics; it will contain terms characteristic of the domain, their definitions, synonyms and translation equivalents, as well as bibliographic descriptions of texts in which they were registered.

It is expected that the Russian thesaurus on corpus linguistics will be integrated into the knowledge portal on computational linguistics developed by a group of Russian researches (Moscow, Novosibirsk and St. Petersburg) [Sokolova et al. 2008].

References

[1] Backer, P.; Hardie, A.; McEnery, T.: A Glossary of Corpus Linguistics. Edinburgh University Press (2006)

[2] Bol'šakova, E. I.; Bajeva, N. V; Bordačenkova, E. A.; Vasiljeva, N. E.; Morozov, S. S.: Leksiko-sintaksičeskije šablony v zadačach avtomatičeskoj obrabotki teksta. In: Kompjuternaja lingvistika i intellektual'nyje tehnologii: Trudy meždunarodnoj konferencii «Dialog–2008». Moscow (2008)

[3] Braslavskij, P. I.; Sokolov, E. A.: Sravnenije p'ati metodov izvlečenija terminov proizvol'noj dliny. In: Kompjuternaja lingvistika i intellektual'nyje tehnologii: Trudy meždunarodnoj konferencii «Dialog–2008». Moscow (2008)

[4] Corpora2002 – Doklady naučnoj konferencii «Korpusnaja lingvistika i lingvističeskije bazy dannych – 2002». St. Petersburg (2002)

[5] Corpora2004, 2006, 2008– Trudy meždunarodnoj konferencii «Korpusnaja lingvistika – 2004 (2006, 2008)». St. Petersburg (2004, 2006, 2008)

[6] Dobrov, B. V.; Lukaševič, N. V.; Syrom'atnikov, S. V.: Formirovanije bazy terminologičeskich sočetanij po tekstam predmetnoj oblasti. In: Trudy 5j Vserossijskoj konferencii «Elektronnyje biblioteki: perspektivnyje metody i technologii, elektronnyje kollekcii» RCDL–2003. St. Petersburg (2003)

[7] Gerd, A. S.: Jazyk nauki i techniki kak object lingvističeskogo izučenija. In: A.S. Gerd. Prikladnaja lingvistika. St. Petersburg (2005)

[8] Kupść, A.: Extraction automatique de termes à partir de textes polonaise. In: TALN 2007. Toulouse (2007)

[9] Levická, J.: Terminology and Terminological Activities in the Present-Day Slovakia. In: Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava (2007)

[10] Mitrofanova, O.; Panicheva, P.; Savitsky V.: Automatic Word Clustering in Russian Texts based on Latent Semantic Analysis. In: Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava (2007)

[11] Rabčevskij, E. A.; Bulatova, G. I.; Šarafutdinov, I. M.: Formalizm zapisi leksiko-sintaksičeskich šablonov v zadače avtomatizacii postrojenija ontologij. In: Trudy 10j Vserossijskoj konferencii «Elektronnyje biblioteki: perspektivnyje metody i technologii, elektronnyje kollekcii» RCDL–2008. Dubna (2008)

[12] Sidorova, E. A.: Podchod k postrojeniju predmetnych slovarej po korpusu tekstov In: Trudy meždunarodnoj konferencii «Korpusnaja lingvistika – 2008». St. Petersburg (2008)

[13] Sokolova, E. G.; Kononenko, I. S.; Zagorulko, Ju. A.: Problemy opisanija kompjuternoj lingvistiki v vide ontologii dl'a portala znanij. In: Kompjuternaja lingvistika i intellektual'nyje tehnologii: Trudy meždunarodnoj konferencii «Dialog–2008». Moscow (2008)

[14] Šajkevič, A. Ja.: Statističeskij slovar' jazyka Dostojevskogo. Vvedenije. Moscow (2003)

[15] Šimková, M.: Výberový slovník termínov z počítačovej a korpusovej lingvistiky (2006) URL: http://korpus.juls.savba.sk/publications/block1/2006-simkova-vyberovy%20slovnik%20terminov/2006-simkova-vyberovy%20slovnik%20terminov.pdf

[16] Urbańska, D.; Piechociński, D.: Automatic Term Recognition in Polish Texts. In: Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava (2007)

[17] Vinogradova, N. V.; Mitrofanova, O. A.: Formal'naja ontologija kak instrument sistematizacii dannych v russkojazyčnom korpuse tekstov po korpusnoj lingvistike. In: Trudy meždunarodnoj konferencii «Korpusnaja lingvistika – 2008». St. Petersburg (2008)

[18] Zakharov, V. P.: Korpusnaja lingvistika / Učebno-metodičeskoje posobije. St. Petersburg (2005)

[19] Zakharov, V. P.; Khokhlova, M. V.: Statističeskij metod vyjavlenija kollokacij. In: Jazykovaja inženerija v poiske smyslov. XI Vserossijskaja objedinennaja konfererncija «Internet i sovremennoje obsčestvo». St. Petersburg (2008)

Using Speech and Handwriting Recognition in Electronic School Worksheets

Marek Nagy

Department of Applied Informatics,
Faculty of Mathematics, Physics and Informatics, Comenius University
Mlynská dolina, 842 48 Bratislava, Slovak Republic
mnagy@ii.fmph.uniba.sk,
http://www.ii.fmph.uniba.sk/~mnagy

Abstract. At elementary schools typical paper worksheets are used. This approach has some disadvantages. A bit clever children fill in it quickly and then they start to bore although they have done some mistakes. From this point of view electronic worksheets with an instant feedback will be more appropriate. There is not problem adopt activities in click-click style but more important activities have writing, reading and listening style at early grades. In this contribution the worksheets for Slovak ABC book Lipka were studied and adopted. To make the instant feedback speech and handwriting recognition is used. An attention will be focused on an isolated word speech recognition/confirmation where a phonetically based approach is used.

1 Introduction

Children use typical paper worksheets at elementary schools. Many kinds of exercises can be depicted on the worksheets. Children read a legend of an exercise and fill in it. This schema is especially problematic for young children who cannot read very well. Of course, first graders cannot read absolutely. This problem is solved by teachers. The teacher explains the exercise and all children of class following him. After, children can start filling in the exercise. It is clearly that this approach needs a synchronization of all children. But a bit clever children fill in it quickly and then they start to bore although they have done some mistakes. The teacher has not time to correct all mistakes of children quickly. Electronic worksheets with an instant feedback can be more appropriate for this purpose.

I will concentrate on first-graders worksheets in this contribution. I analyzed Slovak ABC book Lipka [3] and an appertaining worksheet [2]. The problem with text legend which cannot be read by children is partially solved in [2] by a set of pictographs. These pictographs can help to identify types of exercises and help to manage an adaptation in an electronic form. One exercise from the worksheet is depicted in Figure 1. Typical activities are listen to, read aloud/entitle, encircle, write down or color (Figure 2). I divide these activities into two groups: vocal activities and writing activities. The writing activities use handwriting recognition mainly. Two approaches must be considered. If children write down letters according to patterns then dynamic programming comparison would be used [1, 7]. On the other hand, if children fill in letters or numbers then independent recognition which uses statistical approach is better [11].

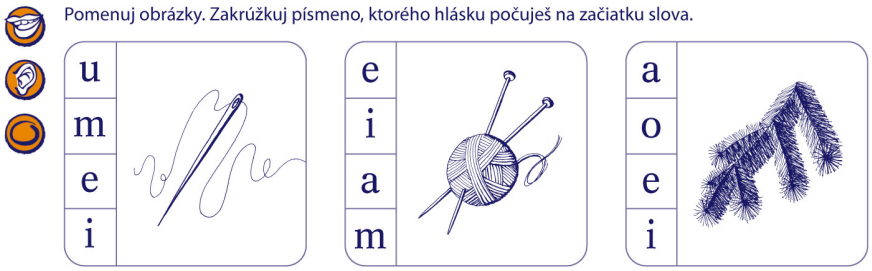


Fig. 1. An example of an typical exercise from worksheet of ABC book Lipka [2]. The legend of exercise in English: "Entitle the pictures. Encircle letters whose phones you can hear at the beginning of words."

Fig. 2. Some pictographs for vocal (the upper row) and writing (the lower row) activities from ABC book Lipka [3].

Next, an attention is given to vocal activities. A typical scenario is a short text which must be read by children aloud. This can be handled same manner as an reading tutor [8]. Children read word by word in microphone and a speech recognizer evaluate accuracy of them. It is not only due to speech recognition accuracy, which is better for isolated words, but the 1st graders learn to read whole isolated words at first. A typical recognition vocabulary includes all words from the story which size is not very big (cca 200) - tight vocabulary. The recognizer uses HTK toolkit [14] and achieves quite good speaker independent word accuracy [8]. Application time responses are less than 1 sec. A recorded isolated word is cut and recognized as a whole. Children can also listen to text which is prerecorded. The recording is forced aligned to text so that children can see an actual spoken highlighted word of the text [9, 10, 13].

It seems that an activity where children entitle pictures can be handled same way. But a problem arose. While the reading tutor can use limited tight recognition vocabulary the entitle activity has to use a big one. To cover all Slovak words with their forms the vocabulary must be sized at hundreds of thousands. An ordinary expansion of the HTK recognizer leads to problems with time responses. To avoid this problem the main vocabulary can be reduced by an approximation of recognized phoneme sequence at the beginning. It will be good if the reduced vocabulary contains a right word. It will give chance the recognizer to choose it.

SAMPA	Slovak word	SAMPA	Slovak word
Vowels			
a	p**a**pier	a:	p**á**s
E	p**e**ro	E:	nov**é**
I	p**i**vo	I:	p**í**skať
O	p**o**pol	O:	p**ó**l
U	p**u**to	U:	p**ú**pava
{	m**ä**so		
Diphthongs			
I_^a	p**ia**tok	I_^E	sp**ie**vať
I_^U/	cudz**iu**	U_^O	k**ôň**
Consonants			
p	po**p**ol	b	ža**b**a
t	va**t**a	d	vo**d**a
c	pla**t**iť	J/	há**ď**a
k	pá**k**a	g	a**g**át
ts	ma**c**o	dz	há**dz**ať
tS	ma**č**atá	dZ	há**dž**em
f	**f**ajka	v	slo**v**o
f_v	**v**dova	U_^	pra**v**da
s	o**s**em	z	vá**z**a
S	košeľa	Z	ve**ž**a
x	**ch**ata	h/	no**h**a
G	vr**ch** hory		
j	dva**j**a	I_^	kra**j**
r	pa**r**a	r=	p**r**st
r=:	v**ŕ**ba		
l	ska**l**a	l=	v**l**k
l=:	v**ĺ**ča	L	**ľ**avý
m	ma**m**a	F	a**m**fiteáter
n	vra**n**a	n	i**n**žinier
N	ce**n**gať	J	va**ň**a

Table 1. Slovak SAMPA symbols with examples

2 Phoneme sequence recognition

Hidden Markov Models (HMM) of Slovak monophones and triphones have been trained during speech recognition training procedure [8]. Parameters of used data and the recognizer accuracy are depicted in Table 2. The baseline recognizer uses tight vocabulary and triphone (context phone) HMM models.

The data are from a project Multimedia reading book [13]. Speakers are children at age 6-12 and patterns were taken in different uncontrolled conditions (different microphones, high SNR, ...). Used phonemes are showed in Table 1 with word examples and were chosen according to Slovak SAMPA [4].

The phonemes recognizer uses one loop grammar of all Slovak phonemes. To improve accuracy a bi-phone (bi-gram) matrix has been computed from training data.

data	speakers	wavs	words	accuracy%
train	352	35934	8293	92.1
test	23	2705	474	86.4

Table 2. The baseline speech recognition accuracy.

	Accuracy %			
	phoneme models		triphone models	
data		with bi-gram		with bi-gram
train	76.9 (80.5)	78.5 (81.8)	95.2 (96.0)	95.6 (96.2)
test	79.8 (84.3)	80.4 (84.9)	88.3 (90.5)	88.1 (90.5)

Table 3. The accuracy of individual phonemes. Numbers in parenthesis show an correction after introduction long and short vowels equivalency.

	Accuracy %			
	phoneme models		triphone models	
data		with bi-gram		with bi-gram
train	12.8 (15.6)	22.0 (26.7)	52.0 (53.4)	65.1 (66.8)
test	4.7 (6.6)	12.2 (16.9)	11.3 (11.9)	21.0 (22.4)

Table 4. The accuracy of whole sequence of phonemes. Numbers in parenthesis show an correction after introduction long and short vowels equivalency.

When the typical triphone training procedure is retraining isolated phoneme models the forced align algorithm is used to determine time boundary of phonemes [14]. The best phoneme sequence is chosen from several ones in phonetics dictionary (HVite command [14]). So transcribed training data are used to compute “bi-gram” statistics matrix. The HLStats command from HTK [14] is used. Then the matrix is included into the recognition “loop” network.

3 Phonetic dictionary searching

The recognized phoneme sequence is converted into word candidates in the next step. Whole phonetic dictionary [5] of 82180 words is searched to choose candidates. The Levenshtein distance algorithm is used for this purpose. But possible operations (insertion, deletion, substitution) are weighted.

3.1 Insertions and deletions

The insertions and deletions are basic operations. It is important which phoneme is actually being deleted. It can happen that vowels are duplicated in the phoneme sequence. It is not so significant for similarity and therefore weight of this deletion or insertion is defined as 1. A special situation is with composite phonemes dz, dZ, ts, tS. The consonants d, t last short time and the phoneme recognizer can only detect the second

part. It is depicted in Table 5 as a similarity (dz ↔ d and dz ↔ z and so on) because it is made as a special substitution rule in an application.

dz ↔ d, z
dZ ↔ d, Z
ts ↔ t, s
tS ↔ t, S

Table 5. Insertion and deletion in composite phonemes. (weight 1)

All vowels, diphthongs and unstopped l=, r= are inserted or deleted with weight 3 because they are base elements of a word (syllable). Other cases have weight 2.

3.2 Substitutions

Sometimes the phoneme recognizer has problems with vowel duration. It has problems to correctly determine which variant of a vowel is better: short or long. And therefore it will be good to consider both variants. Rules are given in Table 6. The starting experiment also shows that accuracy is increasing. See Table 3 and the accuracy in the parenthesis. The duration can not be ignored. For example Slovak words zastávka ↔ zástavka differ only in the vowel duration. Besides vowels there are unstopped consonants r, l too.

a ↔ a:
E, ({) ↔ E:
I ↔ I:
O ↔ O:
U ↔ U:
l, (l=) ↔ l=:
r, (r=) ↔ r=:

Table 6. Substitution variants based on duration. (weight 1)

A special case is phoneme { which is not used in ordinary Slovak language now and therefore it can be substituted by the phoneme E. Due to too little training patterns the phoneme G must be replaced by phoneme x. The phoneme l= is same case. The replacement rules are in Table 7.

According to the book [6] the first two formats of vowels can be plotted such as in Figure 3. A tongue position can approximately be plotted same way as the formants. If a tongue do not take the exact position then the recognizer can make mistakes between neighboring vowels. This aspect is covered by Table 8. The weight of these shifts is 3.

A typical phonemes exchange is based on pair consonants that differ only by presence of voice [5]. See Table 9.

{ ↔ E
G ↔ x
l= ↔ l

Table 7. Ignored variants - replaced phonemes (weight 0)

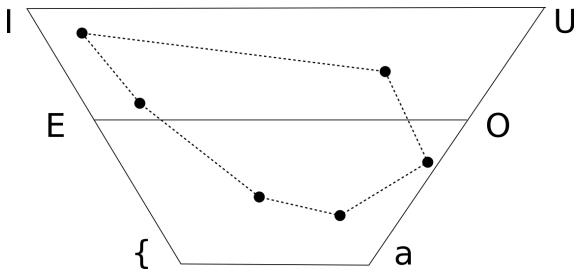


Fig. 3. Vowels continuity diagram (solid line - according to formants, dotted inner line - according to tongue position)

I ↔ E
E ↔ a
a ↔ O
O ↔ U

Table 8. Substitution variants based on vowel shifts (weight 3)

b ↔ p
d ↔ t
J/ ↔ c
g ↔ k
f_v ↔ f
z ↔ s
Z ↔ S
dz ↔ ts
dZ ↔ tS
(G), h/ ↔ x

Table 9. Substitution variants based on pair consonants (voiced/unvoiced) (weight 1)

Table 10 is showing rules based on an assimilation aspect of Slovak language. The assimilation can have a cross words effect. A typical example is u-v assimilation (e.g. strojov). The f-v (e.g. včela) assimilation has own alophone f_v but speakers tend to speak it either as f or as v. Other type is i-j assimilation in diphthongs especially (e.g. piatok, ahoj). The next n-N assimilation rule is based on an inadequate pronouncing. And the last rules are based on non-literary pronouncing (e.g. môj, mój, moj).

U ↔ U_^ ↔ v ↔ f_v
v ↔ f
I ↔ I_^ ↔ j
n ↔ N
m ↔ F
U_^O ↔ O:, O

Table 10. Substitution variants based on assimilation (weight 1)

d ↔ J/	z ↔ Z
t ↔ c	s ↔ S
n ↔ J	dz ↔ dZ
l ↔ L	ts ↔ tS

Table 11. Substitution variants based on palatalization (weight 1)

The palatalization is other aspect of Slovak language. It is marked by a hiccup above consonant graphemes or not if graphemes e, i follow after d, t, n, l. This hiccup omitance makes problems because it can be read hard or floppily. Also some Slovak dialects have tendency read it non-literary hard everywhere and some speakers read it floppily everywhere. The palatalized pairs are collected in Table 11.

A special case is when a composite phoneme is demerged into two parts. It requires to include a new operation in Levenshtein algorithm. The substitution step can go through two symbols but only in one direction (horizontally or vertically). See Figure 4. All cases are depicted in Table 12.

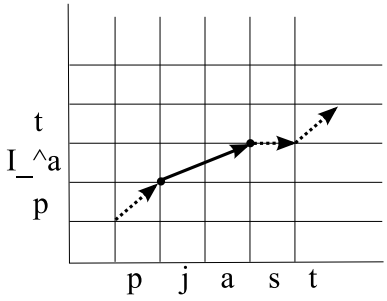


Fig. 4. Example of the new operation (decomposition) of composite phonemes. The arrow can go through two cells either horizontally or vertically.

Two substituted vowels or diphthongs which are not among mentioned rules will be substituted with weight 5. A vowel ↔ diphthong substitution has weight 7. All other not mentioned substitutions have default weight 3.

I_^a ↔	I, I_^, j + a
I_^E ↔	I, I_^, j + E
I_^U/ ↔	I, I_^, j + U
U_^O ↔	U, U_^, v, f_v + O
dz ↔	t, d + dz, z
dZ ↔	t, d + dZ, Z
ts ↔	t, d + ts, s
tS ↔	t, d + tS, S
N ↔	n + g, k

Table 12. Substitution variants based on decomposition of phonemes (weight 1)

4 Results

The new isolated word recognizer is assembled from three stages. At the first, the phonetic sequence is estimated. At the second, the N-best candidates are searched in the phonetic dictionary [12]. And at the third, one best candidate is chosen. The accuracy of the recognition and a time per a word are showed in Table 13. The reached accuracy of the three stage recognizer is comparable with a typical HTK recognizer which uses a beam width search heuristic [14]. Of course it was a goal that the accuracy will not be worse. But time responses are a little better. Three summed numbers in Table 13 represent time consumption of the individual stages. The time reduction achieves 36%. Some time reserves can be seen at the phonetic dictionary searching stage. Now the complete dictionary is saved on a remote MySQL database server. This database server is used by other applications too. The search time can be reduced by rational local storing in computer memory. A little (not mentioned) experiments show reductions close to 1 second. In practical application it will be better connect all three stages together closer and reduce response time to minimum 1 second.

Only two variants of N-best candidates were chosen (50-best, 100-best) for comparison. The 100 candidates are sufficient as can be seen in Table 13. It seems that fixed number of candidates is not so good. Now two words with same distance can be separated one among the candidates and one outside of them. In future it will be better to use distance criteria too. So that all words which have same distance will be altogether among candidates or not.

		through 50-best		through 100-best		basic		basic + beam w.	
data	words	Accuracy	sec/w	Accuracy	sec/w	Accuracy	sec/w	Accuracy	sec/w
test	82180	66.4	3+5+1	67.2	3+5+1	66.1	63	67.6	14

Table 13. The accuracy of the three stage recognition. Triphone models with bi-phone statistics were used. Also the basic recognizer with the beam width heuristic [14] is presented for a comparison.

5 Conclusions

In this paper the problems of electronic school worksheets are analyzed. The goal is to transfer paper worksheet exercises, which are developed by pedagogic professionals, from paper into computers and enrich them by additional functionality. The ABC book Lipka [3] was taken as a basis where exercises are more allowable to this idea. Two types of activities were identified: vocal activities and writing activities. In this article the typical entitle exercises were enriched by the speech recognition approach. The electronic worksheet can check if kids speak correct words that represent the pictures. The isolated word speech recognition is used for this reason. Because kids can speak random Slovak words the recognizer's dictionary must be extended. In this article the whole phonetic dictionary contains 82180 words that are in base forms mostly. The Slovak paper phonetic dictionary has been adopted [12]. This caused increasing the time response. Therefore the better search approach is presented in this contribution. It reduces the time response by more massive vocabulary reduction than the typical Viterbi beam width search approach [14]. The obtained time reduction is significant and with better phonetic dictionary setup it can be close to 1 second per a word in future.

Acknowledgments

This work was partially supported by the Slovak Scientific Grant Agency (VEGA) under the contract No. 1/4060/07 and by the Slovak Research and Development Agency under the contract No. LPP-0056-07.

References

[1] Červeň, J.: Writing tutor – computer helps children learn to write, In Proceedings of the 9th international conference on Informatics, Informatics 2007, SSAKI, Bratislava
[2] Hirková, A., Nemčíková, M.: Pracovný zošit pre 1. ročník základných škôl k šlabikáru Lipka, Aitec, Bratislava 2008
[3] Hirková, A., Nemčíková, M.: Šlabikár Lipka pre 1. ročník základných škôl, Aitec, Bratislava 2008
[4] Ivanecký, J., Nábělková, M.: Fonetická transkripcia SAMPA a slovenčina, Jazykovedný časopis, Vol. 53, No. 2, 2002, Bratislava, pp. 81 – 95
[5] Král', Á.: Pravidlá slovenskej výslovnosti, (tretie vydanie 1996), SPN, Bratislava 1983
[6] Král', Á, Sabol, J.: Fonetika a Fonológia, (prvé vydanie), SPN, Bratislava 1989
[7] Lipková, J.: Počítač ako učiteľ písania (Computer as an writing tutor), master's degree thesis, Comenius University, 2008, Bratislava
[8] Nagy, M.: Čítanie s tučniakom, využitie rečového komunikačného rozhrania pri výučbe čítania detí na I.stupni ZŠ, In Proceedings of the 7th international conference, APLIMAT 2008, Bratislava, Slovakia, pp. 101 – 110
[9] Nagy, M.: Multimediálna čítanka 2008, 4.ročník, Active Advice, Bratislava, 2008, 273 str., ISSN 1337 5601

[10] Nagy, M.: Počítač dokáže naučiť plynulo čítať, Učiteľské noviny, č. 2, 2008, Bratislava, str. 31
[11] Tomacha, M.: Počítač – učiteľ písania (Computer – a writing tutor), master's degree thesis, Comenius University, 2007, Bratislava
[12] Vančo, P., Nagy, M.: Creating of Slovak Electronic Phonetic Dictionary for Use in Speech Recognition, In Proceedings of the 3rd international conference, Slovko 2005, Bratislava, pp. 216 – 219
[13] (www): Projekt Multimediálna čítanka, http://cpr.ii.fmph.uniba.sk/citanka
[14] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.2, Cambridge, England, Cambridge University, 2002

Composite Lexical Units as an Element of Lexicographical Historical Computer System*

Irina Nekipelova

Izhevsk State Technical University, Russia
nekipelova@udm.net

Abstract. This article is devoted to the development in the field of modelling the description of a word lexical value in the information-search system "Manuscript". This aspect is connected with a problem of use a system for realization of linguistic researches in the field of lexics and semantics and working out the linguistic search system allowing the user to have exact idea of a word lexical value and its semantic connections in language and texts of ancient manuscripts, stored in a database.
Keywords: history of Russian language, grammatical semantics, triodion.

1 Introduction

One of the way to save and accumulation information is presentation and research of ancient script monument in special portals. This is a multiple-aspect appearance, an important factor in humanization of social relation and in process of forming of cultural and scientific picture of the world.

Creation of full-text database, forming of corpora of texts Russian early manuscripts and their researching favour the decision of some of important tasks for linguistic science: 1) electronic publishing of ancient Slavonic script monuments favoures the saving of some part of cultural and scientific heritage that keeps in the vaults of libraries in poor condition that's why they don't available for researching often; 2) lingo-textual description of different time ancient lists is a necessary element for study of ancient script monuments as important part of national culture component; 3) different aspect research of lexical composition and semantic relations of language of ancient Russian texts favoures more deep and detailed understanding of history of language and history of the Slavs life; 4) studies in lexical and grammatical semantics favoures the research of peculiarity of functioning language units in ancient texts in different period of language development because language semantics is an important element in forming of the Slavs linguistic picture of the world.

* Работа выполнена в рамках аналитической ведомственной целевой программы Федерального агентства по образованию Российской Федерации "Развитие научного потенциала высшей школы (2009-2010 годы), проект "Лингвотекстологические и корпусные исследования грамматической семантики древнерусского текста" (регистрационный номер 2.1.3/2987).

Electronic publishing of transcription of Slavonic triodion early manuscripts and their lexico-semantic and grammar-semantics peculiarity research presuppose to development and adoption a landmark method of account of vocabulary with a glance of its system and functional properties and words semantic relations. Results of research will have an applied significance for lexicography, lexicology, semantics and history of language because system of the Old Russian language lexical units which use in modern lists of triodions is that of as an adjuvant to Russian language teaching practice and attracting attention of scientists and all comers to that little-investigated manuscripts.

The most promising technique of manuscripts integrated study is electronic collection creating to full-text database. Quality of that collection is in saving information content, fast-access retrieval, full-featured of date sample and sort, possibility of replenishment a database with some new information and visibility of date in the Internet.

Electronic publishing of transcription of Slavonic triodion early manuscripts and their lexico-semantic and grammar-semantics peculiarity research presuppose to development and adoption a landmark method of account of vocabulary with a glance of its system and functional properties and words semantic relations. Results of research will have an applied significance for lexicography, lexicology, semantics and history of language because system of the Old Russian language lexical units which use in modern lists of triodions is that of as an adjuvant to Russian language teaching practice and attracting attention of scientists and all comers to that little-investigated manuscripts. Thereat in information analysis system “Manuscript” (http://manuscript-s.ru/) create databases which include collections of different genres the Slavic script monuments. Collection of triodion creates in their midst.

Nowadays collection includes such publication: Festal Triodion, notificationly, the end of XII century (ГИМ, Воскр. 27 перг.), 206 p.; Lenten Triodion, notificationly, XII century(ГИМ, Син. 319), 315p.; Festal Triodion, services of Twelve Great Feasts and for chosen saints, the beginning of XII century, the end of XII – the beginning of XIII centuries and XIV century(РГАДА, ф. 281 (Син. тип.), № 139), 208p.; Festal Triodion, XIIcentury, РГАДА, ф. 381, № 138),173 p.; Lenten Triodion, the end of XIII – the beginning of XIV centuries (РНБ, Погод. 41), 157 p. Expected expansion of triodions collections within the bounds of intended to electronic publishing such manuscripts: Lenten and Festal Triodion (Triodion by Moses Kiyanin), the end of XII – the beginning of XIII centuries (РГАДА, ф. 381, № 137), 257 p.; Feasts with insertion from Festal Triodion, the second part of XVIII century (РНБ, Соф. 385), 150 p. About importance the text of Triodion by Moses Kiyanin for study of Russian language history wrote A.I. Sobolevskii [1]. Most scientists thought that this manuscript one of the most significant creation in the period of XI-XIV centuries.

Texts of manuscripts must be presented integratly showing general text information in the movement of Slavic culture. So-called virtual libraries give an opportunity

not only to view documents, but receive information about their structure, state, description, etc.

Use of this system makes it possible to get material at short notice for linguistic, historical, philological and lingo-textual research, its full and exact lexico-semantic analysis that give an opportunity a wide range of users IAS “Manuscript” to go beyond the scope of the research of single text and/or contemporary texts and to draw a significant fundamental conclusion on the ground of text collections or corpus of Old and Middle Russian periods.

Electronic publication is necessary to direct usage of gathered scientific matter for linguistic problems solution in the range of lexicon and semantics of the Old Russian language and for linguistic research of triodions lists, i.e. for analysis and modelling lexical matter, which meets in triodion. So, within the bounds of linguistic research of used vocabulary analysis is carried out within the range every triodion and description lexical and semantic the Old Russian word relations that presented in every list of triodion. Lexical description and electronic historical dictionary creation are necessary first of all for modern thinking more understandable of language of manuscript’s text and come across for users of reading. Thus for development of multifunctional web-modules that contained publication of ancient and medieval Slavic manuscripts; developments in the range of lexical meaning of the word and its bounds in the range of semantics modelling are actual.

2 Semantic model

Types of lexical description are foundation of lexical meaning modelling and word’s semantics in database. It should be noted that all meaning types, considered as complementary for each other, i.e. as a constituent (side, aspect, part of all). It’s important because some facts, that complicated research, must be considered in describing word’s semantics in its history. First of all, word has remained to this day only in the certain movement that hampers fixing all possible usages of word and its bounds with other language units, so long as extant texts cannot reflect all word relations, that realized in any period of language’s development. Secondly, in interpreting of bounds of words that use in manuscripts cannot always be used to say about that analysis absolute adequacy, i.e. description of word’s semantics is regarded through the alembic of availability of word’s semantic bounds in modern language which might not be at that lexeme at earlier period of language development.

Nowadays, structure of word’s description was developed that was presented in the form of hierarchical bounds of words. Modelling of semantics is development of typical structure of semantic description, which structure’s usage of capacity is made conditional on individual characteristics and relations of words.

At first, nominative type of sense is worked out by us, because this type represents lexico-semantic bounds of words. Nominative sense is presented with meanings of non-structure lexical units: lexemes and phraseological units. In this case we tell

about lexical and phraseological sense, respectively. For lexical description of word leading is identification of lexical meaning of word, in which seven basic types of description of lexical meaning of word are presented such as encyclopaedic, explanatory, etymological, synonymous, antonymous and homonymous. For phraseological description is important phraseologation of component units in history of language, i.e. their relation to vocal formulas which are spreaded in early stage of language history, or to phraseological units proper, which are appeared at a later period.

Dictionary must contain references to all usage of concerned lexeme in context, i.e. full information about all usage in any context.

Let's present this structure of description by example of word *water*, which was reiterated made of in text of Festal Triodion, XI-XII century (РГАДА, ф. 381 (Син. тип.)), №138.

Nominative sense:

1. lexical sense:

1.1. defining sense:

1.1.1. linguistic sense: *вода*, ж. "вода, естественная влага" [2]. К этому разделу словаря должны быть отсылки всех употреблений слова вода в следующих контекстах: **съ водамн мнѡгамн ндоша къ кладѧзю** (114 л.), **водъı проснлъ ѥсн** (112 л.), **да тамо нъінѣ тъщншнсѧ ежѥ ѡт водъı тѣмь внно сътворн** (039 л.), **нстоуашаго водоу н сухa камъıка** (052 л.), **водоу къ оумъıваннцю въланваше ногъı же оумъıваше родъмь бо владка** (023 об.).

1.1.2. context sense: *бессмертные воды*, *вода бессмертия* – то же, что "живая вода, животная вода": **вода бесъмьртнѧ** (112 л.), **н женѣ просѧщн бесъмьртънъıѧ водъı посъılаешн хвалѧщннмъ мънѡгое твое съмотренне** (116 л.); *вода премудрости*: **вѣрою прншъдъшн къкладѧзоу самарѧнъінн ѩко вндѣ премоудростн водъı напонтнсѧ обнльнѣ н црство въішънѥѥ наслѣдовавъшн вѣуъноѥ ѩко преславьнаѧ** (116 об.), **ѧше н самарѧнъінн женѣ премоудростн водоу таннъствъмь нзлнвъшн** (114 л.), **премоудростн водоу таннъствъмь нзлнвъшн** (114 л.); *воды веры* – то же, что "источник веры": **съ водамн мнѡгамн ндоша къ кладѧзю н ѩко влю нде же бѣ нстоуннкъ прмрстн водъ спсннѧ самарѧнъінн поуърпатн подажъ мн водъı вѣръı да въспрннмоу коупѣлнъıѧ строуѧ** (114 л.); *воды спасения* – то же, что "воды веры": **съ водамн мнѡгамн ндоша къ кладѧзю н ѩко влю нде же бѣ нстоуннкъ прмрстн водъ спсннѧ самарѧнъінн поуърпатн подажъ мн водъı вѣръı да въспрннмоу коупѣлнъıѧ строуѧ** (114 л.); *неповинная вода* – то же что "чистая вода, нежертвенная": **предъ соудншъмъ пнлатовъмь да распѧнѥтсѧ въпнѧхоу всн хе бе оумъıвъін же роуцѣ водою неповнннною кръвью агньуею особѣ осоуженнемь тръсть омуаше** (041 л.);

1.2. encyclopaedic sense: вода (оксид водорода), простейшее устойчивое химическое соединение водорода с кислородом, H2O <...> [3].

1.3. etymological sense: 1) вода, -ы, *ж*. - «бесцветная, более или менее прозрачная жидкость, являющаяся главной составной частью гидросферы земного шара, образующая реки, озера, моря, океаны». *Прил*. водный, -ая, -ое, водяной, -ая, -ое, отсюда водянистый, -ая, -ое. Укр. вода́, во́дний, -а, -е, водяни́й, -а, -е, водяни́стий, -а, -е, во́дявий, -а, -е; блр. вада́, во́дны, -ая, -ае, вадзяны, -а́я, -о́е, вадзяні́сты, -ая, -ае; болг. вода́, во́ден, -дна, -дно, водни́ст, -а, -о, водни́кав, -а, -о; с.-хорв. вòда - «вода», «река», вöднū, -ā, -ō- «водный», «водяной», вöден(ū), -а, -о, - «водяной», «водянистый», вòдньикав(ū) -а, -о — «водянистый»; словен. voda, voden : vodni, -a, -o; чеш. voda, vodácký, -á, -é - «водный», vodnatý, -á, -é – «многоводный», «водянистый»; словац. voda, vodný, -á, -é - «водный», «водяной», vodnatý - «водянистый»; польск. woda, wodny, -a, -e - «водный», «водяной», wodnisty, -a, -e; в.-луж. woda, wódny, -a, -e - «водный», «водяной», wodnawy, -a, -e - «водянистый»; н.-луж. wóda, wódny, -a,-e - «водный», «водяной», wodniaty, wódowaty, -a, -e - «водянистый». Др.-рус. (с XI в.) вода, водьнъ, водьный, значительно позже (с XVI в.) водяный (Срезневский, I, 276, 279; Доп. , 36). Ст-сл. вода, водьнъ, водьныи, водьскъ, -а, -о (SJS. 1:5, 205, 207). ▫ О.-с. *voda. И.-е. база *aued- : *ŭd-; им. ед. *u̯édor : *u̯ódō(r) |> о.-с. *voda]; ср. локатив ед. *udén(i), род. ед. udnés (см. Pokorny, I, 78). На славянской почве родственные образовании: о.-с. *vydra (см. *выдра*), о.-с. *vědro (см. *ведро*). Ср. гот. watō - «вода»; то же др.-в.-нем. waszar (совр. нем. Wasser); англосакс. wœter (совр. англ. water); греч. *ϋδωρ*, фригийск. *βεδυ* <*uedō); др.-инд. udán(i), локатив ед., udnáh;, род. ед. от udaká-m - «вода»; хетт. wātar, род. wetenas; с назализованным вокализмом: лит. vanduõ, род. vandeñs (при жем. unduõ), латыш. ūdens - «вода»; др.-прус, unds - тж.; латин. unda - «волна» [4]; 2) вода́, сюда же *во́дка*, укр., блр. *вода́*, др.-русск., ст.-слав. вода *ϋδωρ* (Супр.), болг. *вода́*, сербохорв. вòда, словен. vóda, чеш. voda, слвц. voda, польск. woda, в.-луж., н.-луж. woda. Древние ступени чередования представлены в в*едро́, выдра*. ‖ Родственно лит. vanduõ, род. п. vandeñs, жем. unduo, д.-в.-н. waszar «вода», гот. watō, греч. *ϋδωρ*, ϋδατος, арм. get «река», фриг. βέδυ, др.-инд. udaká́m, uda-, udán- «вода», unátti «бить ключом», «орошать», ōdman- ср. р. «поток», алб. uj «вода»; носовой согласный в лат. unda «волна» и лит. vanduõ вторичного происхождения; см. Вальде 850; И. Шмидт, Pluralb. 202 и сл.; М.-Э. 4, 404 и сл.; Хюбшман 434; Уленбек, Aind. Wb. 28 и сл. Древняя основа на r/n [5]; Вода. Общеслав. индоевр. характера. Та же основа, но с перегласовкой, содержится в словах ведро, выдра [6]. Etymological sense is presented in compliance with information from etymological dictionaries.

1.4. homonymous sense: *вода* – "водное пространство (о море, реке, озере и т. п.)": ѩко тебе на водахъ повѣшъшаговоу ѯемлю недержимо тварь вндѣвъшн на лъбьнѣмъ внсѧщѧ оужасьно одьржима бѣ нѣсть ста развѣ тебе гн въѯываюцѣн (049 л.), водъı морьскъıѧ улѣклбуе бълнъı жнтнискъıнѧ погроужаютъ мѧ (111 об.), дьньсь внснть на древѣ нже на водахъ ѯемлю обѣшнн вѣньць шт тьрнна въскладаетсѧ (037 л.).

1.5. synonymous sense: *вода* (в значении "море") = *пучина*: шъствовалъ есн по поуунѣ немокрыми стопами въплътнсѧ насъ ради (112 л.).

1.6. antonymous sense: *вода* (в значении "море") – *суша*, *земля*: покрываѧ водами превъспрьнѧѧ своѧ полагаѧ н морю предѣлъ пѣтъкъ съдьржан землю (041 л.), расѣдаетсѧ каменнıе н землѧ трѧсетсѧ (035 об.).

2. phraseological sense:

2.1. vocal formulas: *животная вода*: вндѣвшн тѧ жена самарѧнъınн троуднвъшасѧ щедрын проснвъшн пнтн водъı бесъмьртьнъıѧ прнне водоу въпнющн дажъ мн жнвотноую водоу гн да ннколн же жажоу въ вѣкъı дѣ блгте (115 об.), нісе бе мон водою жнвотною помазавъ.хожаше троужаѧсѧ плътню н оуклада зѧсѣдаше (114 л.), н водоу жнвотноу шт тебе прнѧтъ (114 об.), дъвоıе зълосътворн пьрвѣньць снъ моніль ме не оставн нстоуннка водъı жнвотнъı ннскона себѣ кладазь съкроушенъ мене на дрѣвѣ распѧ варавоу же н отъпоустн оуднвнсѧ нбо осемь н снце лоуча съкръı (043 об.); *вода бессмертия*: вода бесъмьртнѧ (112 л.); *бессмертные воды*: н женѣ просящн бесъмьртънъıѧ водъı посълаешн хвалящннмъ мъншгое твое съмотренне (116 л.); *живая вода*: водоу жнвоу нісе обѣщавъ нмъ (112 л.), вода жнва текущн (112 л.); *жизненная вода*: самарѧнъınн женѣ проснвъшн водоу жнзньноу подасть внемь разоумъмь (115 л.).

2.2. phraseological units which aren't presented in text of triodion

Let's suppose that user is interested in meaning of word *вода* and its lexico-semantics bounds with other words in the following movement: да тамо нъınѣ тъщншнсѧ еже шт водъı тѣмь внно сътворн (039 л.). In that case lexeme is referred us to part of dictionary, where information about using of that lexeme is contained. In this case, information presents like that:

Nominative sense:

1. lexical sense:

1.1. defining sense:

1.1.1. linguistic sense: *вода*, ж. "вода, естественная влага": да тамо нъınѣ тъщншнсѧ еже шт водъı тѣмь внно сътворн (039 л.),

1.1.2. context sense isn't presented.

1.2. encyclopaedic sense: вода (оксид водорода), простейшее устойчивое химическое соединение водорода с кислородом, H2O <...>.

1.3. etymological sense: 1) вода, -ы, *ж*. – «бесцветная, более или менее прозрачная жидкость, являющаяся главной составной частью гидросферы земного шара, образующая реки, озера, моря, океаны» <...>; 2) вод*а́*, сюда же *во́дка*, укр., блр. *вода́*, др.-русск., ст.-слав. вода *ϋδωϱ* <...>; 3) вода. Общеслав. индоевр. характера <...>.

1.4. synonymous sense isn't presented

1.5. antonymous sense isn't presented

1.6. homonymous sense: *вода* – "водное пространство (о море, реке, озере и т. п.)": ꙗко тебе на водахъ повѣшъшаговоу землю недержимо тварь вндѣвъшн на лъбьнѣмъ внсѧща оужасьно одьржима бѣ нѣсть ста развѣ тебе гн възъıвающн (049 л.), водъı морьскъıѧ улѣклбуе вълнъı жнтнискъıнѧ погроужаютъ мѧ (111 об.), дьнысь внснть на древѣ нже на водахъ землю обѣшнн вѣньць шт тьрнна въскладаетсѧ (037 л.).

2. phraseological sense:

2.1. vocal formulas isn't presented

2.2. phraseological units isn't presented

However, user may to be interested in meaning of lexeme *вода* (*water*), which homonymous by *water* word in meaning "limpid liquid" and its lexico-semantics bounds with other words in the following movement: ꙗко тебе на водахъ повѣшъшаговоу землю недержимо тварь вндѣвъшн на лъбьнѣмъ внсѧща оужасьно одьржима бѣ нѣсть ста развѣ тебе гн възъıвающн (049 л.). In this case, user receives the following information:

Nominative sense:

1. lexical sense:

1.1. defining sense:

1.1.1. linguistic sense: *вода* – "водное пространство (о море, реке, озере и т. п.)": ꙗко тебе на водахъ повѣшъшаговоу землю недержимо тварь вндѣвъшн на лъбьнѣмъ внсѧща оужасьно одьржима бѣ нѣсть ста развѣ тебе гн възъıвающн (049 л.)

1.1.2. context sense isn't presented

1.2. encyclopaedic sense isn't presented

1.3. etymological sense isn't presented

1.4. homonymous sense: *вода* - "вода, естественная влага": съ водамн мнѡгамн ндоша къ кладѧзю (114 л.), водъı просналъ ѥсн (112 л.), да тамо нъınѣ тъщншнсѧ еже шт водъı тѣмь внно сътворн (039 л.), нстоуашаго водоу н суха камъıка (052 л.), водоу въ оумъıваалннцю въланваше ногъı же оумъıваше родъмь бо влдка (023 об.).

1.5. synonymous sense: *вода* = *пучина*: шъствовалъ есн по поууннѣ немокрънмн стопамн въплътнсѧ насъ радн (112 л.).

1.6. antonymous sense: *вода* (= *море*) – *суша*, *земля*: покрънваѧ водамн превъıспрьнѧѧ своѧ полагаѧ н морю предѣлъ пѣтъкъ съдьржан землю (041 л), расѣдаетсѧ каменнѥ н землѧ трасетсѧ (035 об.).

2. phraseological sense:

2.1. vocal formulas isn't presented

2.2. phraseological units isn't presented

Such placement of lexico-semantics information about word is required certain comments.

Encyclopaedic and etymological interpretations are reflected original relations of word. Encyclopaedic sense is given for reference value of this word. All homonymous lexemes haven't encyclopaedic sense because this type of dictionaries hasn't senses of homonym. The same situation with etymological description of word – derivative homonyms haven't etymological characteristic. So, we see in diverse terms of descriptions and senses of word which don't contradict to each other.

Explanatory sense is presented in the form of the linguistic and context sense. Linguistic sense of word is sense, as a rule, that is fixed in Dictionary of Russian language of XII-XVII centuries and agreed with etymological sense. Senses, that isn't fixed in dictionaries, is described as context senses of significance lexical units with semantic characteristics which they have in reduced context directly.

It's important under interpretation of word to identification of context senses of word. Basic criteria of differentiation of linguistic and context senses of word, consecutive usage of them to description of word functioning reveals linguistic or context character of word meaning, are worked out by us.

Special role in development of word's semantics plays its relations with other words, including stability usage of the word in the define movement.

Semantics of the Old Russian language researching presupposes researching of the important component – grammatical semantics. Grammatical semantics is the integral component of semantics in whole which enables to supplement integral linguistic picture of the world that is reflected in texts of Old and Middle Russian periods.

Grammatical semantics must be seen as separate bloc of textual usage of language unit's researches in ancient scripts monuments that binds semantics and grammar in whole. It is very important under functioning of composite language units in manuscripts.

Grammatical semantics of vocal formulas and phraseological units are attracted particular interest because the define role in development of word's semantics plays relations with other words including stability unstability (lexical and grammatical) usage in the movement, degree of unification of components formation.

Grammatical semantics of set expressions regards of forming of meaning and composition of set expressions and phraseological units. Generally in that period scientifics fix functioning of vocal formulas. Compound lexical units are special lexicographical units that have special characteristics. In whole vocal formulas permit to transposition, to substitution, to admission of components with saving of initial sense, even so grammatical variativity used of components is met often, which may lead to change of semantics.

Description of phraseological sense of word is hard task; because many factors must be take into account such as time of appearance of the formation, field of usage and conversion degree, degree of activity and degree of repeatability in works of different genres, expansion narrowing of phraseological sense and degree of unification of the forming components. It task becomes more complex with respect to the Old Russian language. As is generally known, process of phraseologization has long his-

tory and that set expressions, which are met in texts of scripts monuments of XI-XIV centuries, aren't phraseological units. Generally, in that period scientists don't fix of functioning of vocal formulas.

V. Deryagin reasonable notes: "Для периода сложившейся деловой письменности в языковом отношении под *формулой* целесообразно понимать фразеологизм номинативного или коммуникативного характера, а также словосочетание, синтаксическую конструкцию (модель предложения) с более или менее постоянным лексическим составом. В отдельных случаях формула может состоять из нескольких предложений, связанных между собой синтаксически и по смыслу" [7]. Formula is a basic unit of stylistic analysis of official text. It is unit of text level and with it formula may be define in terms that use for units in other levels that lower in hierarchy: formula is sentence (definite type), formula is set expression (definite type), a phraseological unit [8].

Well, it should be affirmed that availability of linguistic formulas not only in official written language texts, but in text of different genres, because vocal formulas using is defined not only with characteristics of genre, but with general linguistic processes. One of the means of forming of vocal formulas is semantic tracing of Greek metaphors, which is led to its symbolization (look in formings **бесмертнаѧ вода**, **непобѣдимаѧ вода** and so forth). Term *formula* V.V. Kolesov correlates with wording of a borrowed symbols of Greek culture in Old Russian texts. "Древнейшие [заимствования] не были свободны от контекстов, в составе которых они перешли к славянам, и эти контексты попадали к ним в письменной форме переведенных текстов. Заимствовалось словосочетание целиком, почему и заимствованные слова оказывались фразеологически связанными" [9].

In the capacity of working-class definition in our research we take which suggests N.A. Antadze. Formula is a special set expression; scientist attributes them to phraseological combinations by common feature of invariability common meaning and relating usage in examples. Formulas admit of shift, substitution, passing of components with preservation of invariable common meaning [10]. So, often formulas are set expressions that are bound syntactically (in the context) and phraseologically (by sense, content), seldom formulas are sentences which are characterized by stability and repeatability [11].

Structure of detection of phraseological possibility of word **водѧ**demands some comments. In semantic model "vocal formulas" field always fills. It should be noted that lexical units which are marked in that field not always have fixed in different dictionary sense; in field of "phraseological units" is empty in terms of usage of capacity by citations from text of triodion, but in that field dates are showed in phraseological dictionary, which reflects well-established phraseology of present language. It is natural that this two fields do not agree and haven't common date.

Fields, where phraseological sense is fixed, fill as far as possible bulking of research materials.

In the capacity of material of research of grammatical semantics wordforms are taken, such elements of composite units that met in text of Festal Triodion XII century (РГАДА, ф. 381, №138, 173 л.). Compound units of language with stable lexical-syntactic character with a component **вода** had been revealed in the Festal Triodion:

Животная вода in the meaning of "well, running water, drinkable water" [Словарь 1975, 2: 249]: **нісє бе мон водою жнвотною помазавъ. хожаше троужаасѧ плътню н оуклада ѧ засѣдаше** (114 л.); **н водоу жнвотноу шт тебє прнатъ** (114 об.); **вндѣвшн тѧ жена самарѧнъінн троуднвъшасѧ щедрын проснвъшн пнтн водъі бесъмьртьнъіѧ прнне водоу въпнющн дажъ мн жнвотноую водоу гн да ннколн же жажоу въ вѣкъі дѣ блгтє** (115 об.); **дъвоє зълосътворн пьрвѣньць снъ моніль ме не оставн нстоуннка водъі жнвотнъі ннскона себѣ кладазь съкроушенъ мене на дрѣвѣ распѧ варавоу же н отъпоустн оудневнсѧ нбо осемь н снце лоуча съкрън** (043 об.);

Живая вода in the meaning of "well, running water, drinkable water" [Словарь 1975, 2: 249]: **водоу жнвоу нісє обѣщавъ нмъ** (112 л.); **вода жнва текущн** (112 л.);

Жизненная вода the same as *живая вода/животная вода* in the meaning of "drinkable water" [Словарь 1975, 2: 249]: **самарѧнъінн женѣ проснвъшн водоу жнзньноу подасть бнемь разоумъмь** (115 л.);

Бессмертные воды the same as the живая вода/животная вода in the meaning of "well, running water, drinkable water": **н женѣ просѧщн бесъмьртънъіѧ водъі посълаєшн хвалѧщннмъ мънщгое твое съмотренне** (116 л.);

Вода бессмертия in the meaning of "drinkable water": **вода бесъмьртнѧ** (112 л.).

In whole we say only about two vocal formulas which have lexico-grammar variations:

1. We say about functioning of vocal formula *живая вода/животная вода/жизненная вода* in the meaning of "well, running water, drinkable water" which composed of only the first component is changed: *живая, животная, жизненная.* Meaning of that component goes back to meaning of word *жизнь*. It assigns meaning of all compound units – "water that needs for life support". «Единство составного имени основывалось на единстве обозначаемого им понятия. Грамматически подчиненным компонентом являлось прилагательное, а в качестве семантически подчиненного компонента выступало существительное, поскольку оно определяло лишь частичную принадлежность образования... Прилагательные несли основную семантическую нагрузку» [12]. It should be noted that substitution of vocal formula's component doesn't change its meaning. All formula's lexical variations are used to in similar and identical contest. It should be said that variation *жизненная* вода is used more infrequent.

Grammatical variation, that is observed in functioning of this compound units of language in the text, is expressed: 1) reverse order of components in composed of for-

mula, for example, *вода животная*: нісє бе мон водою жівотною помазавъ.хожаше троужаѧсѧ плътню и оуклада ѕасѣдаше (114 л.); и водоу жівотноу шт тебе прнатъ (114 об.) preorder *животная вода*: въпнюшн даждъ мн жівотноую водоу гн да ннколн же жажоу въ вѣкъı дѣ блгте (115 об.); 2) using of adjective forms that don't segment instead of segment forms, for example, *вода животна/воды животны*: не оставн нстоуннка водъı жівотнъı ннскона себѣ кладаѕь съкроушенъ мене на дрѣвѣ распѧ баравоу же и отъпоустн оуднвнсѧ нво осемь и снце лоууа съкръı (043 об.), well as *вода жива*: водоу жівоу нісє обѣщавъ нмъ (112 л.); вода жнва текушн (112 л.) and *вода жизньна*: самарѧнъınн женѣ проснвъшн водоу жнѕньноу подасть внемь раѕоумъмь (115 л.).

2. The next vocal formula, that we find in the text of Festal Triodion is *бессмертные воды*. Meaning of that language unit didn't fix in historical dictionary. This meaning associated with definite of drinking water, but it isn't become formed with word *жизнь*, but with *смерть*. In semantic point the last formula is oppose to the formula *живая/жизненная/животная вода*. If the second formula has object meaning "drinking water", then the first formula *бессмертная вода* has in the text a symbolic significance, i.e. "not only for life supporting, but for deliverance from death, acquisition of eternal life". That semantic difference with difference of lexical components, i.e. we say about two different vocal formulas, not only one. Lexico-syntactic meaning will be more and more parted when be used often.

Grammatical variation composed of that formula is presented with syntactic relation's change of components of compound element. Noun construction immortal water with agreement bound with main noun and dependent adjective: и женѣ просѧщн бесъмьртънъıѧ водъı посълаешн хвалѧщнмъ мъншгое твое съмотренне (116 л.); вода бесъмьртна (112 л.). It substitutes construction *вода бессмертия* with control bound with main noun and dependent noun: вода бесъмьртна (112 л.). It is important that grammatical change compound of that formula don't reduce to semantic change in the structure of word meaning.

Analysis of functioning of vocal formulas in texts of old written records allowed to state a fact of usage incomplete structure. But observable vocal formulas are characterized by only full structure. This is due to the fact that loss compound of leading in semantic point of adjective leads down to disappearance of the formation, i.e. it is been unrecognizable in the text.

In whole the process of phraseologization of compound language units is characterized by long history. In many cases vocal formulas were become phraseologies and had more abstract meanings with the lapse of time. But part of vocal formulas, which were used in text of XI-XIV centuries, were lost and been absent in nowadays language system. All formation that were researched don't fixed in phraseological dictionaries. Only *живая вода* is fixed and as a result it has meanings: Живая вода 1. Фольк. Мифическая чудодейственная жидкость, возвращающая жизнь

мёртвому телу. 2. Экспрес. Всё, что одухотворяет, благотворно действует, пробуждает интерес [13].

Modern meanings of phraseology *живая вода* are bound with reference value. The object meaning and the subsequent symbolical meaning giving rise to that phraseologycalized the abstract meaning. It should be noted that lexical and grammatical variations of our phraseology don't observe nowadays. We see that in phraseology dictionaries.

Grammatico-semantic analysis of compound language units functioning in old ancient script texts is important for reconstruction of relations and bounds of words in the Old Russian. Therewith the analysis is important for modern lexicography system development, because it shows the permanent development of language system. It reflects specific character of ancient Slavs language world picture.

3 Conclusion

It should be noted, that all content words have all presented characteristics, signs and bounds. But only some of that words maximal realize this signs. Analysis of lexical definitions mustn't overshadow a real vocal using, contest and realizable lexical units functions, actualizations of expansion or contraction a conception and forming of context meanings, i.e. somewhat that composes a variable part of language.

Necessary part of research is detection of lexemes that are used in written records. Lexemes and their meanings didn't fix in Old Russian dictionaries. It is important not only for triodions description as independent genre, but for changing of conception about manuscripts are created in XII-XIII centuries.

In the issue of all word analysis that usage in triodions must be create hierarchical lexico-semantic language system model; lexicography field that included all bounds and word relations of Old Russian language too, because totally of meanings are formed semantic language system, i.e. system of meanings.

References

[1] Соболевский А.И. Труды по истории русского языка. Т.1.: Очерки из истории русского языка. Лекции по истории русского языка / Предисл. и коммент. В.Б. Крысько. – М.: Языки славянской культуры, 2004. – С. 12.

[2] Словарь русского языка XI-XVII вв. В 17-ти т. Т.2. – М.: Наука, 1975. – С. 249.

[3] Большая Российская энциклопедия: В 30 т. Т.5. Великий князь – Восходящий узел орбиты. – М.: Большая Российская энциклопедия, 2006. – С. 481-483.

[4] Черных П.Я. Историко-этимологический словарь современного русского языка: В 2 т. / П.Я. Черных. – М.: Рус. яз., 1999. Т.1.: А-Пантомима. – С. 159.

[5] Фасмер М. Этимологический словарь русского языка. Т.1. (А-Д) / М. Фасмер. – М.: Прогресс, 1986. – С. 330.
[6] Шанский Н.М., Краткий этимологический словарь русского языка / Н.М. Шанский, В.В. Иванов, Т.В. Шанская. – М.: Просвещение, 1971. – С. 85.
[7] Дерягин В.Я. Варьирование в формулах русской деловой речи XV-XVII вв. / В.Я. Дерягин // Восточные славяне: Языки. История. Культура: К 85-летию академика В.И. Борковского. – М.: Наука, 1985. – С. 243.
[8] Дерягин В.Я. Варьирование в формулах русской деловой речи XV-XVII вв. / В.Я. Дерягин // Восточные славяне: Языки. История. Культура: К 85-летию академика В.И. Борковского. – М.: Наука, 1985. – С. 243.
[9] Колесов В.В. Философия русского слова / В.В. Колесов. – СПб.: ЮНА, 2002. – С. 201.
[10] Антадзе Н.А. Лексико-фразеологический состав Судебников 1497 и 1550 гг.: Автореферат диссертации на соискание ученой степени кандидата филологических наук / Н.А. Антадзе; Тбил. гос. ун-т. - Тбилиси, 1965. – С.16–18.
[11] Некипелова И.М. Метонимическая и метафорическая деривация в истории русского языка (на материале памятников деловой письменности XI-XVII веков). Автореферат диссертации на соискание ученой степени кандидата филологических наук. – Казань, 2005. – С. 4.
[12] Некипелова И.М. Метонимическая и метафорическая деривация в истории русского языка (на материале памятников деловой письменности XI-XVII веков). Автореферат диссертации на соискание ученой степени кандидата филологических наук. – Казань, 2005. – С.11–12.
[13] Фразеологический словарь русского литературного языка: В 2 т. / Сост. А.И. Фёдоров. Т.1.: А-М. – М.: Цитадель, 1997. – С. 87–88.

IT: Moving Towards Real Multilingualism

Antoni Oliver and Cristina Borrell

Universitat Oberta de Catalunya, Spain

Abstract. The Linguamón-UOC Chair in Multilingualism of the Universitat Oberta de Catalunya has developed a project consisting of the automatic elaboration of linguistic resources, those including Catalan. For this, we have created an automatic extractor of terminology, which is freely distributed, multi-platform and adaptable to users' needs. One of its most important useful applications is the elaboration of glossaries, both monolingual and multilingual, based on a set of documents. The system automatically extracts the term candidates from the texts introduced by the user. Later, a specialist reviews the list of lexical units to verify the result. In the Chair in Multilingualism we have applied this tool to the expansion of the Eurovoc glossary with its Catalan version. We have extracted the equivalents for this language from the entries in Spanish and a bilingual Spanish – Catalan parallel corpus. From the application of the extractor, we obtained a multilingual glossary of 2 531 terms involving EU policies. In a second stage, we have used different techniques. Firstly, starting from the terms that were not yet translated by the first step, we have generated hypothesis of the Catalan translation of the Spanish Eurovoc terms, thus by the compositionality method. Secondly, we have applied statistically automatic translation, and finally we have checked the results in a monolingual corpus or by means of a restricted Web search. Our aim is to translate automatically as much terms as possible. With this project we try to demonstrate that having the suitable resources can remarkably reduce the translation costs.

1 Introduction

Multilingualism is one of the most visible aspects of the change society is living at present. A new challenge is growing, namely to cope with this change in any sphere of our lives. In this framework, the major aim of the Linguamón-UOC Chair in Multilingualism of the Universitat Oberta de Catalunya is promoting a concept of linguistic diversity that is sustainable, equitable and functional. For that reason, our team has developed several projects aiming to highlight the importance of linguistic richness and thus give a chance to all languages, even minor ones.

In this regard, the Linguamón-UOC Chair in Multilingualism has developed a set of open source tools aiming to automatically extract terminology. Our team has applied these tools to develop Catalan terminology resources related to the European Union. Currently, Catalan has an ambiguous status within the EU and very few documents are translated to this language. One of the reasons that are adduced not to translate them is the assumption of high costs and the consideration of Catalan as a minority language. With this project we try to demonstrate that having the suitable resources and using assisted and automatic translation tools, translation costs are

remarkably reduced. Moreover, with a modest budget, the right of Catalan citizens to have the documentation in their own language can be respected.

As we all know, the European Union produces official documents in 23 official languages[1]. The primary law documents must be written in all EU official languages, in order to respect the transparency, democracy and legitimacy that the Community promotes. Despite this, many Europeans encounter limitations when it comes to participating in the European institutions, as much of the multilingual and multicultural richness that exists within the borders of the EU is not taken into account.

In order to manage the volume of multilingual documents and assist both citizens and community workers, the European Union has set up a big amount of specialized language services[2]. For instance, the translation service in the European Commission is the most representative EU linguistic service, and one of the largest translation services in the world.

The European Commission's Directorate-General for Translation (DGT) has about 2,500 workers (thus apart from freelance translators) and produces about 1.3 million pages per year. The translation units are organized into departments, one for each EU language. In recent years, the translation process has been refined, especially with the incorporation of the technology offered by translation memories, so that translators can optimize their effort. Roughly, each translator uses to develop his or her work:

- Terminology (dictionaries, glossaries, terminological databases, etc.).
- Parallel and reference documents;
- Access to previously translated texts (especially through translation memories);
- Administration staff responsible for fulfilling the tasks of pre-edition and post-edition.

Another interesting aspect of the DGT is known as "field offices for multilingualism", which in some way, are responsible for promoting the richness of the linguistic diversity of Europe. The function of these branches is to adapt the Commission's communications to the languages of the region and act as a bridge between the local citizens and European institutions.

Thanks to the bilateral administrative arrangements between the EU institutions and the Spanish government[3], since June 13th 2005, Catalan citizens may apply to the institutions in their own language. They may do it in the written communications to the Council of the European Union and the European Commission, through an intermediate body designed by the state government in charge of translating the

[1] Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish.

[2] All information concerning the languages of Europe and the different language services, both internal and 'open', is available at the website of the EU "Languages and Europe".

[3] Council conclusions on the official use of other languages (2005 / C 148/01).

messages (from Catalan to Spanish), and send the translation to the institution concerned.

2 Tools

Eurovoc is a multilingual thesaurus covering the areas of activity of the European Union. Currently, it is available in Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish. It is the only resource that can be downloaded from its website.

Following this information, our team is able to develop a multilingual glossary with the tools described above. First of all, the team has selected the "available" languages for the project. English is the most popular language in the EU, followed by relatively far behind by French and German. The most logical thing then is to choose English (due to the amount of original documents available) and French, because a romance language will always be closer to Catalan, our target language, rather than German.

We will also take Spanish into consideration. This is one of the closest languages to Catalan, and above all, it is quite easy to find parallel documents in this linguistic combination (Spanish-Catalan). For instance, and even within the institutional field, the Official Journal of the Government of Catalonia[4] is published both in Catalan and Spanish.

All that linguistic potential can not be exploited, combined and adapted if we do not have any tool that helps us handle. That is why we have programmed and used an automatic extractor of terminology[5] that allows us to obtain the equivalent in Catalan from the Spanish Eurovoc terms. Basically, we have given the Spanish Eurovoc terms to the official parallel corpus in its Spanish version, thus receiving in return the official, Catalan version of the term.

Thanks to this collection of tools we can automatically select, manipulate and retrieve terminology, which is of great help for the translator. It is distributed under a free software license (GNU/GPL), so the user can freely download it, use it, distribute it or modify it following his needs. The tools are programmed in Perl, in a completely modular way, in order to provide, if necessary, its modification and adaptation. The package works for different platforms, for both Linux and Windows.

This tool allows users to:

- extract term candidates directly from the text;
- get help in finding the best translation for a word;
- automatically build glossaries, both monolingual and multilingual;
- create terminological databases for specific fields;
- automatically build lists of lexical units of a specific area.

If the input of the system is a monolingual corpus, the resulting output is a monolingual list of term candidates. This list must be reviewed manually by a

[4] DOGC `http://www.gencat.net/dogc/`

[5] Downloads and more information `http://ww.linguoc.cat`

linguist, who must confirm or reject the proposals automatically selected by the extractor. Otherwise, if the input consists of a set of documents in multiple languages, apart from the list we obtain a multilingual terminological resource, because the system automatically detects any possible equivalent translation of a candidate.

This tool kit is essentially a linguistic terminology extractor and a device for automatic translation equivalents search, programmed essentially with statistical methods.

1. The automatic terminological extractor takes a document, a set of documents or a parallel corpus and stop−words lists (lists of functional words) of the languages involved. The statistical method identifies term candidates based on word frequency in a corpus of expertise. The process, broadly speaking, is this: the system calculates statistically n-grams (sequence of n elements, usually from n = 2 to an n determined by the user, i.e. all combinations of two words, three words etc.). Logically, some of these combinations will not be relevant, that is why it is essential to filter the results with the list of empty words. The output of this module will already be the list of term candidates.
2. In the second case, the system picks a term candidate in a language A and selects a subset of the parallel corpus with all segments containing this term. If, then, the system makes a statistical terminological extraction in these segments, the result should be the term candidate in the language B. It is expected that the terminological unit more common in the segment in the language B would be the term that we searched in language A.

3 Procedure

By using the Eurovoc entries in Spanish (6.797 items) and the official, bilingual Spanish – Catalan parallel corpus we have extracted the equivalents in this language. The input corpus selected, the Official Journal of the Government of Catalonia (Diari Oficial de la Generalitat de Catalunya, from now on DOGC), is published daily from Monday to Friday (except holidays) in two equivalent editions, one in Catalan and another in Spanish. Basically, the laws, rules and regulations, general dispositions, agreements, resolutions, edicts, notices, ads and other acts of Government and its Administration are published in it, both in Catalan and Spanish.

Thanks to it, we have a parallel Catalan-Spanish corpus of institutional, political and legal areas, that it is why it is one of the Catalan official texts which shares more characteristics with the EU institutional documents. At the same time, as it offers a Spanish version, it optimizes the work of the extractor: it simply searches the Spanish Eurovoc terms in the Spanish version of the DOGC and returns the Catalan translation for it.

Although the statistical method resolves successfully the search in many cases, it is not always right. So, the program offers more than one candidate, leaving the user the final choice. The first result is always the one that has appeared more frequently.

Here is where the work really starts for the linguist. Now he or she has to manually review all the proposals made by the extractor, correct them when they are incorrect and even suggest translations if the extractor has been unable to offer any. This will happen in the cases where the Spanish word we are searching for does not appear in the DOGC. Either way, thanks to the extractor, the linguist saves time and research.

If the extractor did not act satisfactorily, the expert reviewing the outcome will have to deal with the types of problems mentioned here[6]:

1. Often, the first proposal for the Catalan translation is not complete. This happens especially in cases of terminological units consisting of more than one word, which somehow mislead statistical systems.
2. Sometimes, Catalan and Spanish versions are not equivalent. This happens not often, but it can be that the extractor mistakes and confuses the selection of terms that do not belong to the same field of expertise.
3. Sometimes the extractor proposes n-grams belonging to the same segment as the key term, but that do not correspond to the possible translation.
4. It can be that versions in other languages other than Spanish do not correspond to that of Catalan. It is recalled that the extractor takes the word in Spanish to make the choice of Catalan, so it never takes into account other languages (in this case, English and French)[7].
5. It can also happen that the version in Catalan is the development of an abbreviation or vice versa. The expert will have to take that into account and harmonize criteria.

In a second stage, we will translate the terms use different techniques. Firstly, starting from the terms that are not yet translated, we will generate hypothesis of the Catalan translation of the Spanish Eurovoc terms, thus by the compositionality method. Secondly, we will apply statistically automatic translation, and finally we will check the results in a monolingual corpus or by means of a restricted Web search. Our aim is to translate automatically as much terms as possible.

4 Results

To perform the experiments we used a parallel corpus Spanish-Catalan (the DOGC) with all issues from 3.544 (dated 2/01/2002) to 5.118 (dated 24/04/2008). The corpus has been segmented by sentences and automatically aligned. It consists of a total of 9 492 333 segments and 121 145 821 words in both languages.

[6] These errors were found after the implementation of the first version of the extractor, reason why most of which disappear due to the improvements in later versions. This analysis helped us discover what changes were important to improve the performance of the extractor.

[7] Ultimately, this would depend on the quality of the Eurovoc equivalents, which is outside the scope of this study.

The results presented were first obtained for an increment value of n ± 1 and the search algorithm of translation equivalents was not optimized. Thus the first candidate to appear was the one that showed the greatest frequency of occurrence in the parallel subcorpus of all the segments containing the original term. The table shows the percentage of equivalents appearing in first position, in the first three and in the top ten positions.

Position	**Percentage**
1	43,4
1 – 3	73,4
1 – 10	86,2

Table 1. Results obtained with the algorithm without optimization.

Analysing the results we reached the following conclusions:

- When candidates get into the top positions with the same frequency, the candidate who comes in first place can be any of them. This is because hashes are used to store the candidates, and these data structures do not have a certain order.
- Often the first candidate has a higher frequency than the original word. This means that mono−word terms will appear in the first place even if the original term was multi−word.

To minimize the consequences of the first phenomenon we programmed an optimization of the translations' equivalent search algorithm which consisted in gathering all the proposals that had the highest frequency and apply the algorithm on the edit distance (the closer the candidate is to the original, the more likely to be the correct one it is), the results improved as follows:

Position	**Percentage**
1	62,2
1 - 3	82,1
1 - 10	88,2

Table 2. Results obtained grouping the more frequent candidates and using the edit distance.

To avoid the problem produced by candidates with a fake higher frequency we have applied an optimization similar to the one above, but that including all three candidates with the highest frequencies. Thus we obtain the following results:

Position	Percentage
1	74,7
1 - 3	80,4
1 - 10	88,2

Table 3. Results obtained by pooling the three candidates with the higher frequencies and using the edit distance.

As we can see, the results have improved markedly, especially in the top positions.

Results of the second stage will appear during the second half of the year.

5 Conclusions and future work

In the development of the project we have obtained a multilingual glossary of 2 364 terms used in EU official documents, available in four languages, among which Catalan. For the moment, above all the official language resources it has only been possible to process the Eurovoc, which is the only tool that offers the possibility to be directly downloaded.

During the development we could improve the algorithm three times, in order to obtain better results. The final result is that the extractor has a 75% of success in the very first position. The success percentage raises to 88.2% when we look at the first ten positions: that means that within the first ten choices the translator can find the correct translation candidate. From the application of the extractor, we obtained a multilingual glossary of 2 364 terms involving EU policies.

In a second stage, we have used different techniques. Firstly, starting from the terms that are not yet translated (6 797 – 2 364 = 4 433 terms left), we have generated hypothesis of the Catalan translation of the Spanish Eurovoc terms, thus by the compositionality method. Secondly, we have applied statistically automatic translation, and finally we have checked the results in a monolingual corpus or by means of a restricted Web search. Our aim is to translate automatically as many terms as possible.

In sum, we have shown that some basic and simple tools can help us create multilingual resources that can expedite the process of translating documents. With the right tools, the process of generation of these resources is very flexible and, therefore, economic. In addition, all tools used are freely distributed and therefore available for the whole community[8].

This methodology can be applied to a large number of languages. The only requirements are namely to have a parallel corpus of the desired areas of expertise and a short list of empty words of the pairs of languages involved. One should keep in mind that the statistical methodology for the extraction of terminology does not work well for agglutinative languages.

[8] Download and more information about the tools: http://multilingualismchair.uoc.edu.

Future work is divided into two lines:

- On the one hand, to apply the methodology described to other European terminology resources, such as IATE.
- On the other hand, to continue working on improving and optimizing the extraction tools.

References

[1] European Comission. (2008). Un reto provechoso. *Cómo la multiplicidad de lenguas podría contribuír a la consolidación de Europa* [on line]. Brussels. Internet: http://ec.europa.eu/education/languages/archive/doc/maalouf/report_es.pdf

[2] Oliver, A., Vazquez, M. & More, J. (2007). Linguoc LexTerm: una herramienta de extracción automática de terminología gratuita (Tomo 11) [on line]. *Poughkeepsie: Translation Journal*, number 4. ISSN 1536-7207. Internet: http://accurapid.com/journal/42linguoc.htm

[3] European Comission. (2004). Translation and drafting aids in the European Union languages [on line]. Brussels. Internet: http://ec.europa.eu/translation/index_en.htm

[4] Diari Oficial de la Generalitat de Catalunya (DOGC) [on line]. Barcelona: Generalitat de Catalunya, 2008. Internet: http://www.gencat.net/dogc/

[5] Eurovoc. Multilingual Thesaurus [on line]. Brussels: European Union, 2005. Internet: http://europa.eu/eurovoc/

Introduction of Non-Verbal Means of Communication in the Corpus of Live Speech*

Tatyana Petrova and Olga Lys

Far Eastern National University, Vladivostok, Russia

Abstract. 1. An integral component of the national corpus is live speech. It causes some complexity when it is introduced into the corpus. In particular, a meta-textual annotation which reflects non-verbal components of communication – con-situation and pair-linguistic communicative means such as gestures, mimics, etc. It is very important to take into consideration the significance of the non-verbal components for understanding oral texts.
2. Remarks reflecting con-situation are necessary in two types of discourse – in speech which accompanies actions and in speech which comments events.
3. Remarks for pair-linguistic means, which substitute verbal means of communication are necessary as they are the carriers of information.

Anthropocentrism of new linguistics has caused "change in theoretical priorities"- first of all, interest to the discourse, reflecting the whole process of real communication. According to V. A. Plungyan, in the context of a new ideology in linguistics corpus "has become a powerful instrument of analyzing facts of the language", which "has returned to the linguists their real object – texts in a natural language in a maximum volume" [Плунгян 2008]. An integral part of any national corpus is live oral speech, which makes information about a given language complete. It has been recognized "exclusively actual" for Russian language to create a collection of oral texts as the corpus researches start on the material of Russian.

Traditionally by live speech is meant an oral speech used spontaneously in informal communication. It is a special sort of speech – improvisation, created in the process of the communicative act. It causes some complexity when introducing the texts of live speech into the corpus. In particular, a meta-textual annotation which reflects non-verbal components of communication – pair-linguistic communicative means such as gestures, mimics, looks, activity of the partners. Creators of the first text-book of conversational Russian [Русская разговорная речь 1978] brought to notice that sort of difficulty and said that to get an adequate notion of a real conversation, it is necessary to study oral speech in its interaction with non-verbal components which form communicative act. They also mentioned the necessity of expressing oral communication in a written form with maximum accuracy. Non-verbal components play an important part in it: in colloquial linguistics they are con-situation and pair-linguistic means (gestures, mimics, etc.)

* Researches are made with the support of Russian Humanitarian Scientific Fund (Project "Creation of the electronic database "Live speech of far-easterners" (№ 08-04-12103в)

In the National corpus of Russian the phenomena of this kind are expressed by remarks which in opinion of E. A. Grishina "do not hide any intrigue in themselves": they are placed between two "unique marks"(e.g. #All are silent#) and under standard processing they are automatically transferred to another level of intra-textual existence and this fact lets not consider them when statistic analyses is made [Гришина 2005]. A problem should be admitted with the degree of fulfilling the text with such remarks. The same problem arises when a multimedia corpus is created (scripts of films and cartoons, TV- and radio commercials). E.A. Grishina writes: "If there are no strict limits, it will bring to the fact when the text of the creators of the corpus will exceed the text of the film" [Гришина 2005a: 242]. To solve the problem of introducing non-verbal components of communication in the sub-corpus of live speech, in our opinion, it is necessary we take into consideration the data of colloquial linguistics - how significant are such components for understanding an oral discourse.

Remarks serving to introduce pair-linguistic means and remarks reflecting con-situation should be separated. Con-situation is partly demonstrated in meta-description: e.g. The text is characterized by "the place and the time of the events described". However the remarks, reflecting con-situational circumstances, which arise in the process of communication and define the content of the text, are very important. Speech turns out very con-situational if the topic of it is caused by con-situation. Such type of a discourse is a mold of verbal and non-verbal and speech that accompanies action and speech that comments events refer to it first of all. [Земская 2004]. Examples:

Situation 1: On a commuter train an elderly lady (Grandmother@) and a kid (Alina@)

Alina@ #Moves her hair-brush along the seat# This is a boat floating. #She drops her ball; a young man picks it up#

Grandmother@ What should you say?

Alina@ Thank you. #She tries to bandage her knee and asks her grandmother# Bandage! #Grandmother bandages Alina's knee but she tries to unbind granny's scarf# Granny, unbind! #Grandmother is silent.# Unbind! #Grandmother is silent but starts to unbind# Where is my blood coming from? That's it; I am going to bind up. Where are my socks?

Grandmother@ #Takes her socks# Put them on!

Alina@ #Looks into the bag# What is in there?

Remarks in the given fragment are necessary since speech, accompanying the actions of communicators, is only an additional component of their activity and the purpose of interaction is non-verbal actions.

Situation 2: On a commuter train, women are returning home from the country.

Woman1@ # She sees a jumper# A jumper is coming.

Woman2@ I am about to get off.

Woman3@ #Is looking out of the window# Oh! Second River, it's getting cooler already.

Woman1@ Second River – what a smell! How people are living here!

Woman2@ #Lookes out of the window# Some pipes.

Woman1@ What pipes! Effluent from all over the city!

Woman3 @ #Lookes out of the window# Fishermen are catching fish.

Woman1@ A then sell that fish, contaminated with helminthes.

The content of the text is determined only by the con-situation – the view from the window of the commuter train. In this case the use of the remark (#Look out of the window#) might be possible: the circumstance determines the discourse.

Before saying about introducing pair-linguistic means of communication in the corpus, it is necessary to explain what it is. The use of mimics, gestures and other pair-linguistic means is a peculiar feature of a chat. Such means are used "not only to accompany, to illustrate or enhance the speech but as information carrier as well - substitution of a verbal component" [Земская 2004]. As G. E. Kreydlin says, "in the act of conversation different systems of processing sign information co-exist and they present a complex proportion" [Крейдлин 2004]. It should be said about a various use of the term "gesture": in a narrow notion it is just a gesture, i.e. movement with hands, with a head, with legs; in a wide notion it is also a movement with a body, postures, expression of the face. Apart from gesticulation all these gestures are the elements of a communicative act.

All said above lets say about necessity of introducing remarks reflecting pair-linguistic elements of an oral discourse. In our opinion, when organizing the corpus of live speech, two types of gesture remarks must be differentiated - obligatory and optional. For this purpose the function of pair-linguistic means should be taken into consideration with respect to verbal communication. As it is known, there are two groups of gestures in colloquial linguistics: "sign" – with both a plan of expression and a plan of content and "non-sign" – with a plan of expression only: they express rhythm or emotional shadow of an expression [Русская разговорная речь 1983]. Remarks, reflecting rhythmic or emotional gestures, may be considered optional not obligatory as they don't have sign function – so important for understanding of a text. Thus, remarks of the type #Bows his head#; Well, I don't know! #He makes a helpless gesture"; How could I forget! #She covers her eyes with hands# and so on may become "unnecessary" for the corpus. As a rule, such gestures double a verbal sign so it is possible to get along without them.

"Gesture-signs" have a plan of content and transmit information, they can be demonstrative, descriptive and symbolic [Русская разговорная речь 1983]. The function of substitution is very significant because when it is fulfilled, a non-verbal means (gesture, mimics, action) become a carrier of information and substitute a verbal component of communication. Gestures of all three types are various and closely interact with live speech. Nevertheless, it depends how necessary is to use the remarks, reflecting sign gestures, since a gesture and content of speech interact in various ways in the boundaries of notional expression.

Gestures-expressions which can be both a motivation and a reaction in a dialogue should become mandatory elements of short-hand notes introduced into the corpus of an oral text. Example:

Situation 3:

Woman1@ He is from Artem, isn't he?

Woman2@ #She shrugs her shoulders#

The example above includes symbolic gesture "I don't know" and introduction of such remark is necessary for keeping the wholeness of the discourse.

Situation 4:

At the seaside, two women – Mother@ and Daughter@

Daughter@ #Goes into the water. Shrinks with cold#

Mother@ What? The water is cold? It's not time to swim.

Daughter@ Oh, no! It's just from the beginning, later it will be OK.

In a given fragment a symbolic gesture with a meaning "cold" has a function of a phrase-motivation which organizes minimal dialogue integrity. However, gestures from the examples above could accompany verbal expressions as well (e.g. #Goes into the water. Shrinks# Cold! Or #Shrugs his shoulders# I don't know). In this case, gesture frames would be optional, so might be omitted in short-hand notes of the oral text.

Remarks, denoting demonstrative gestures, and which are tied closely with demonstrative pronouns might be necessary or mandatory in the discourse in a various degree. Examples:

Situation 5:

Two relatives are in the room – Aunt@ and Niece@.

Auth@ #She thumbs the recipe-book# One more thing. I'll write down the juice of grapes. And I'll make it later.

Niece@ Here is the paper. #She points to the desk# From the wild?

Aunt@ #Bows her head to the cat# Lying. Hypnotizing. Yes, from the wild.

This fragment vividly demonstrates sense integrity of verbal and non-verbal components of communication that is very characteristic of an oral discourse. So the necessity of the remarks, reflecting both con-situation and demonstrative gestures is obvious. But gesture remarks of that kind are not obligatory in the cases when a gesture just accentuates a demonstrative pronoun while accompanying it. Example:

Situation 6:

Woman1@ Mushrooms are sold everywhere.

Woman2@ Yes, we'been walking along today. [#Points to the pan with mushrooms#] That's all we've gathered.

Situation 7:

Junior members of the family (Husband@ and Wife@) tell their grandmother:

Wife@ We went to the beach yesterday. Heaps of crabs there were there!

Husband@ Heaps! They are soft, faded completely. They are so soft!

Wife@ They are so small. [#Demonstrates#]Only one was a palm size. So… So many crabs! #Makes a gesture "up to the neck"#

In the past phrase it is a relevant remark and it demonstrates a symbolic gesture meaning "a lot".

Remarks denoting figurative gestures are not always necessary. They might be omitted in case the meaning of the gesture is also expressed verbally. Example:

Situation 8:

Two relatives are in the room –Aunt@ and Niece@

Niece@ How is Irishka doing at school?

Aunt@ It's been just the second week she is at school. She took the same English as yours, like this. [#Shows with her hands#] Sort of a great book.
Niece@ Ah, by Drozdova.

Thus, in the corpus of live speech obligatory are the remarks to express sign pair-linguistic means which substitute a verbal component of communication and become a carrier of information. In future an approximate list of the remarks on the basis of typology developed in the colloquial linguistics might be made.

Solving the problem of introducing non-verbal means of communication in the corpus of live speech is important not only in respect of "orthographic principle of representing oral speech". The system of the remarks, set by definite rules, will be necessary in case of making the sound-track to the short-hand reports possible. No doubt, a transcript plays an important part, i.e. is how accurately con-situational circumstances and non-verbal means of communication are expressed, as without them the whole discourse can't be understood properly. And it is also obvious that for linguistic research the corpus of live speech may become complete only when it reflects the picture of real communication in situations of a chat, where we can't get along without a non-verbal component.

References

[1] Гришина Е. А. О принципах размещения устных текстов в Национальном корпусе русского языка // НТИ. Сер. 2. – 2005. – № 3. – С. 31 – 38.
[2] Гришина Е. А. Два новых проекта для национального корпуса: мультимедийный подкорпус и подкорпус названий // Национальный корпус русского языка: 2003 – 2005. Результаты и перспективы. – М.: Индрик, 2005а. – С. 233 – 250.
[3] Земская Е. А. Разновидности городской устной речи и задачи их изучения // Земская Е. А. Язык как деятельность: Морфема. Слово. Речь. – М.: Языки славянской культуры, 2004. – С. 247 – 289.
[4] Крейдлин Г. Е. Невербальная семиотика: Язык тела и естественный язык. – М.: Новое литературное обозрение, 2004.
[5] Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. – М.: Языки славянской культуры, 2008. – С. 7 – 20.
[6] Русская разговорная речь. Тексты / Отв. ред. Е. А. Земская и Л. А. Капанадзе. – М.: Наука, 1978.
[7] Русская разговорная речь: Фонетика. Морфология. Лексика. Жест / Отв. ред. Е. А. Земская. – М.: Наука, 1983.

MorphCon – A Software for Conversion of Czech Morphological Tagsets

Petr Pořízka[1] and Markus Schäfer[2]

[1]Department of Czech Studies, Faculty of Arts,
Palacký University in Olomouc, Czech Republic
petr.porizka@upol.cz
[2]Institute of Computer Science
The University of Bonn, Germany
schaefe1@informatik.uni-bonn.de

Abstract. This study reflects current situation in Czech corpus linguistics with a special view to morphological annotation of language corpora. Several morphological tagsets of Czech exist nowadays. These tagsets differ by the conception reflecting morphological categories in different extent of complexity. There has also been no possibility of conversion among tagsets. New tool called *MorphCon* (Morphological Convertor) is now being developed for these purposes. This first version (0.1alpha) enables converting of two basic morphological tagsets of Czech: Prague positional system and Brno's attributive system. There are three basic Input/Output (I/O) formats of data (SimpleTag-Conversion, KWIC/Tag-Format, WPL-Format) within version 0.1alpha. Tagsets are implemented into the *MorphCon* as "drivers" with "encode" and "decode" function as well as an "universal library" called *DZ-Interset* (© Daniel Zeman) – modified in our tool – plays key role for the process of conversion as a transcoder. The *MorphCon* software is thus built as an universal converter: modularity, the *Interset* as a transcoder, possibility of adding of another tagsets (not only Czech ones) and I/O formats.

1 Introduction

The development of the *MorphCon* software (shortened from ***Morph****ological* ***Con****vertor*), an application based on formal rules and algorithms, was motivated by the current situation of Czech corpus linguistics with respect to morphological annotation of linguistic corpora. At present several morphological tagsets exist for Czech, which reflect morphological categories to various degrees of complexity and which differ one from another in their conceptions. The predominant and most used system is the one designed by J. Hajič (a positional tagset, hereafter, PT) [6][7], as seen in the written part of the *Czech National Corpus* [18]. Not less important is the Brno system of morphological tags (an attributive tagset, hereafter, AT), the conception of which is authored by K. Osolsobě [11] and R. Sedláček [17]. This system is made use of by the morphological analyzer (tagger) called AJKA [15][16] in the corpora of the *Natural Language Processing Centre* at Masaryk University, Brno (hereafter, NLP FI MU).

Among others belongs Petkevič´s morphological tagset used in the international project MULTEXT-EAST (the Orwell 1984 Corpus) [13] or, most recently, the so-called *kódovník* (coder), used lately for tagging of the *Prague Spoken Corpus* [4].

The tagsets mutually differ with respect to their conception, the degree of complexity and system as far as the coded grammatical categories are concerned. Morphological tagsets also depend on particular software applications, which are not composed to annotate linguistic data by means of another tagset, as they operate with one particular tagset only. However, if necessary, it is possible to annotate the same text by various tagsets by means of the respective applications, but the crucial role is played here by the efficacy, i.e. the degree of success and failure at (semi)automatic processing of the texts by the given taggers, the scope and structuring of their dictionary for lemmatization and tagging etc.

One of the possible solutions to this situation is a software allowing automatic conversion of (Czech) morphological tagging systems. For these reasons, the *MorphCon* converter has been developed, which allows the conversion of corpus data already tagged by one tagset into a different tagset.

The conversion between the Prague and Brno tagsets has been chosen for the first phase of this software. The Prague system is a positional one (15 positions are given, to each of which a particular linguistic category is assigned, represented by a defined subset of symbols), the Brno system is an attributive one (there is a combination of diagrams, in which the first symbol represents a grammatical category, the second one its concrete value for the word given). A detailed account and an overview of the symbols of both systems are presented in Hajič [6][7], Osolsobě [11], and Sedláček [15][16][17]. For us to find out the possibilities of the mutual automatic conversion, it was necessary to carry out an analysis and comparison of both tagsets, the key factors being mutual compatibility, loss of information and other aspects (see below).

2 The *MorphCon* converter

The software *MorphCon* (v0.1alpha) is in development since 2008 by a team of authors consisting of members of three universities: Petr Pořízka (Faculty of Arts, Palacký University, Olomouc), Marek Schäfer (Faculty of Informatics, Bonn University, Germany) and Daniel Zeman (Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague). It is developed in the *Perl* scripting language (v5.10.0) and designed as a universal converter, the fundamental elements of which are modularity and existence of a conversional tagset (which excludes the direct conversion). This allows implementation of further tagsets and Input/Output formats.

With respect to possible future users-linguists, the so-called Graphic User Interface (GUI) has been kept in mind.

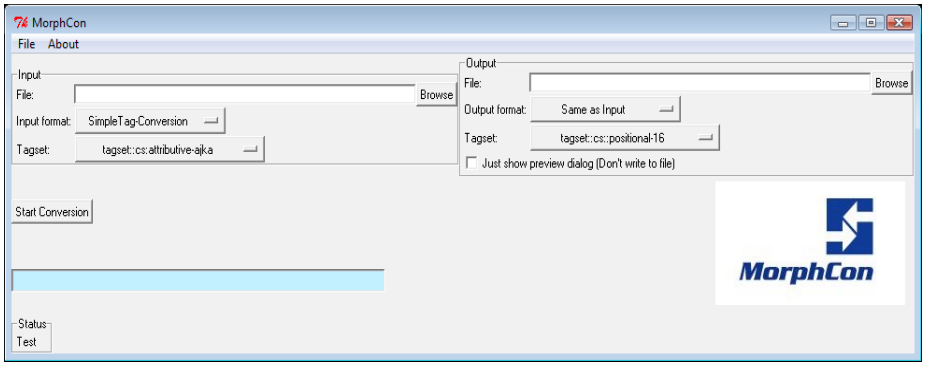


Fig.1. The Graphic User Interface of the MorphCon converter

2.1 The structure of the software

The conception, structure and the principles of the *MorphCon* converter are presented in the *Scheme* in the Supplement, to which the following text refers to. The *MorphCon* is composed of several components/modules (all written in the Perl scripting language):

- GUI: the Graphic User Interface – MorphCon.pl
- Input/Output modules: MorphCon::{simple,kwic,wpl}
- Drivers: implemented morphological tagsets – tagset::cs::{attributive-ajka,pdt,positional-16}
- Universal library: a modified *DZ Interset* library

The key quality is modularity of the software, apart from offering universality (including the possibility to enrich the *MorphCon* with other tagsets) also variability in the process of conversion, i.e. various/different settings of the input and output data. The *MorphCon* is based on the universal tagset *DZ Interset* (see [21] [22]) which works as a transducer when converting one tagset into another. Each tagset is implemented into the software as a „driver“ with double function, either as a source, or target tagset:

- encode-function: source-tagset → Interset
- decode-function: Interset → target-tagset

2.1.1 The Interset

The *Interset* functions as a „feature-projection“, i.e. it is structured as a „feature::value“ system, the grammatical, morphosyntactic category being the feature. It must also contain all the features (grammatical categories) with their values from all implemented tagsets. An overview of all features and drivers may be found on the websites of the project [24]. During the conversion, the tags from the tagset A

are converted into the *Interset*, from which they are subsequently converted into the target tags within the tagset B. The quality of conversion thus depends on the quality of algorithms converting the particular categories with values of the tagsets given into the "feature::value" system of the *DZ Interset* [23].

Example: The *Interset* and the "feature::value" structure for the attributive tag 'k2eAgInPc6d1' from the ajka-tagset in the Perl script:

```
{
 'animateness' => 'inan',
 'case' => 'loc',
 'degree' => 'pos',
 'gender' => 'masc',
 'negativeness' => 'pos',
 'number' => 'plu',
 'pos' => 'adj',
 'possgender' => 'masc',
 'possnumber' => 'plu',
 'tagset' => 'cs::ajka'
}
```

	Annotation of the tag '**k2eAgInPc6d1**'	
k2	k = part of speech	2 = adjective
eA	e = negation	A = affirmation
gI	g = gender	I = masculine inanimate
nP	n = number	P = plural
c6	c = case	6 = sixth
d1	d = degree	1 = positive

The principle of conversion of this tag from an attributive one into a positional tagset with representation in the Perl script in the "AT,PT {Interset} format" looks as follows:

Example:

```
k2eAgInPc6d1,AAIP6MP--1A-----{'animateness' => 'inan','case'
=> 'loc','degree' => 'pos','gender' => 'masc','negativeness'
=> 'pos','number' => 'plu','pos' => 'adj','possgender' =>
'masc','possnumber' => 'plu','tagset' => 'cs::attributive-
ajka'}
```

2.1.2 The input/output modules

The Input/Output modules of the *MorphCon* noticeably extend the options of the *DZ Interset*, since – apart from the process of conversion itself – they allow to set the input and output data variably. Thus, the input and output formats do not have to be identical.

The options for the Input/Output modules

- **the data file**
  - the *.txt* format, "plain text"
- **the tag format**
  - the Brno tagset – attributive: tagset::cs::attributive-ajka
  - the Prague tagset – positional: tagset::cs::pdt
    tagset::cs::positional-16
- **the file format**
  - simple: SimpleTag-Conversion
  - KWIC: KWIC/Tag-Format
  - WPL: WPL-Format

Note on the data file

So far, the *MorphCon* has only been working with the plain text (*.txt* files), but in the future it should also process texts with structured data format (SGML and especially XML data).

Note on the tag format

There are two variants for the positional tagset: (a) "tagset::cs::pdt" is the original driver of the *Interset*, in which the category of aspect is not included; (b) "tagset::cs::positional-16" comes from the original driver, but the important category of aspect has been newly added to it in the *MorphCon*. (This means an extension from 15 to 16 tagset positions.) Therefore the positional system of tags may be enriched by means of conversion in the direction AT → PT, as the attributive system counts with the aspect category, while the positional system does not (apart from some exceptions). It is a way to prevent a loss of information during conversion in such a case when the source tagset contains a grammatical category not found in the target tagset.

Note on the file format

As mentioned above, the *MorphCon* allows the change of input and output file formats. When converting tagset A into tagset B, it is thus possible to determine the resulting data format, differing from the source data format A. So far the *MorphCon* has been working with three formats: (1) The SimpleTag-Conversion (shortened: simple), (2) The KWIC/Tag-Format (shortened: KWIC), and (3) The WPL-Format (shortened: WPL). See the following examples of these formats:

2.1.2.1 The SimpleTag-Conversion

This format is a simple conversion from tagset A into tagset B. One separate line corresponds to one tag of the respective tagset:

AT	PT
k3gInPc6	PDXP6----------
k2eAgInPc6d1	AAFP6----1A-----
k1gInPc6	NNIP6-----A-----

Table 1. The simple "tag-to-tag" conversion

2.1.2.2 The KWIC/Tag-Format

The KWIC/Tag-Format allows the conversion of tags from already annotated and completed corpora. In practice, the so-called concordances are most frequently searched for, i.e. search results where the key word (KWIC = Key Word In Context) is surrounded by a context (see the example taken from the *Bonito* corpus manager – for more detailed information about the *Bonito* concordancer see Rychlý [14]). Usually the tag is placed as a metatext following the word, this means exactly in the format KWIC/Tag (the slash is a division mark). In the converting process, only tags are converted, the remaining text is kept unchanged.

The source tagset (AT) – the pre-conversion phase

172 měl zajištěnou nominaci na Světový pohár v < Petrohradu /k1gInSc6wS > . " Má Švanda naději startovat

284 prezident Václav Havel za přítomnosti dalších < hostí /k1gMnPc2wK > . Za zvuku státní hymny vztyčili

↓

The target tagset (PT) – the post-conversion phase

172 měl zajištěnou nominaci na Světový pohár v < Petrohradu /NNIS2-----A----- > . " Má Švanda naději startovat

284 prezident Václav Havel za přítomnosti dalších < hostí /NNMP2-----A---2- > . Za zvuku státní hymny vztyčili

2.1.2.3 The WPL-Format

The corpus data are usually structured into the so-called vertical format, there is always only one word (WPL = Word Per Line) with its linguistic interpretation per line. For that reason, the sequence *word – lemma – tag* is most often separated by a tabulator (comma or other marks are acceptable, too). The WPL-Format format is therefore very important in case we build a corpus: we annotate the data with the help of tagset A, but we may subsequently need to convert the data into tagset B. The reason for this may be situations when a tagger with implemented tagset A has at its disposal a higher-quality and larger dictionary and gives higher-quality results (i.e. proportionally fewer errors during the process of automatic annotation) than a tagger with implemented tagset B. In comparison with tagset A, tagset B is more user-friendly, may be remembered more easily etc. There may be a whole variety of reasons for using this format.

AT: word	lemma	tag	PT: word	lemma	tag
V	v	k7	V	v	RR--6----------
těch	ten	k3gInPc6	těch	ten	PDXP6----------
dlouhých	dlouhý	k2eAgInPc6d1	dlouhých	dlouhý	AAFP6----1A-----
rozhovorech	rozhovor	k1gInPc6	rozhovorech	rozhovor	NNIP6-----A-----

Table 2. The sequence word – lemma – tag in the WPL-Format

3 Different conceptions of AT and PT systems

Important facts for conversion, i.e. conversion algorithms, are the question of different conceptions of tagsets and their mutual convertibility, the potential loss of information during conversion, and other aspects. There may be cases where modifications of tagsets are concerned, see the example of a positional tagset supplied with another position representing the category of aspect.

Problems when converting the attributive and positional tagsets of Czech appeared, in fact, we still find ourselves at the stage of testing and adjusting the conversion algorithms to optimize the *MorphCon* conversion process.

The positional tagset is characterized by the controversial second position (SUBPOS), which is a rather unsystematic and heterogeneous mixture of symbols from various grammatical categories of various parts of speech and with various degrees of complexity (cf. [7]). The second position contains 75 values altogether, which according to our opinion could be reduced, avoiding loss of relevant linguistic information. The „double-category" of the 3rd/6th and 4th/7th positions (gender and number), and existence of the 6th and 7th categories as separate positions expressing possessive gender and possessive number, respectively, are also debatable. On the other hand, the grammatically important category of aspect is missing. The category of aspect has been indeed added as the 16th position to the annotated corpora of Czech called *SYN2005* and *SYN2006PUB* (cf. [2][3]), but it has not been implemented into the positional tagset (and tagger). The question remains, what the still unexploited „reserved" positions 13 and 14 are intended for, and why the tagset still contains these two empty positions. The attributive tagset is more complex concerning the contained linguistic (sub)categories and could be characterized as the one that is more systematic regarding the conception, more user-friendly regarding acquisition and more easily remembered. The problem that appeared when working on the conversion algorithms concerned documentation and a complete set of tags. There exist several versions of the attributive tagset from various development stages, but there is no version history available, and therefore we had to take all of them into account. There exist four versions of the attributive tagset in total: (1) the official version by R. Sedláček (the author of *Ajka*, the tagger using the attributive tagset), available on the website of NLP FI MU [17]; (2) the tagset table by R. Sedláček from 2006; (3) the tagset version found in the disambiguation manual by Bartůšková et al. [1]; (4) the original tagset version from R. Sedláček´s diploma thesis [15].

A number of Perl scripts were created for testing of the *MorphCon*, which had partial functions during the conversion process, e.g. indicating tags of the same content and visualization of differences between "encode" and "decode" functions of "drivers", i.e. tagsets. With help of testing scripts, the problematic situations and errors originating before the conversion, namely already during the process of annotation of corpus data, may be detected. Mainly non-uniformed tag order, erroneous annotations of particular tokens, which remained in the text most probably due to insufficient disambiguation or were inserted into the tags by a disambiguator (the „human factor“), e.g. an attribute without a value or a value without an attribute, more values in the category of case and other errors.

4 Further advancement of the *MorphCon*

In the near future, we plan to expand functionality of the *MorphCon*. We intend to implement other tagsets of Czech (the above mentioned *mte-cz tagset* or *kódovník*), further Input/Output formats (both linear and vertical), e.g. the format used by the *Ajka* tagger [15] or in the *Prague Dependency Treebank* (the so-called csts /pml/ format – for more details, see *PDT 2.0 Guide*, Chapter 3. Data) [12].

Following the experience acquired during the development of the *MorphCon* and solving problems relating to the conversion, implementation of some scripts originally intended for testing purposes only were considered. The aim is to produce a new module „Tag-Checker“, which would serve as a helping instrument for disambiguation and tag checking (checking their correctness/error rate, anomalies etc.). Basic information about the *MorphCon* software may be found at http://www.morphcon.webnode.cz.

References

[1] Bartůšková, D., Hlaváčková, D., Ungermanová, M. (2004). *Manuál pro značkování a desambiguaci slovních tvarů v jazykových korpusech*. Brno. [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/projekty/desman/desman1603.pdf>

[2] *Corpus SYN2005*. [online], [cit. 2009-05-29]. Available from: <http://www.korpus.cz/english/syn2005.php>

[3] *Corpus SYN2006PUB*. [online], [cit. 2009-05-29]. Available from: <http://www.korpus.cz/english/syn2006pub.php>

[4] Čermák, F. et al. (2007). *Frekvenční slovník mluvené češtiny*. Praha: Karolinum.

[5] Hajič, J., Hladká, B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset, In: *Proceedings of the Conference COLING – ACL `98*. Montreal, Canada, pp. 483–490.

[6] Hajič, J. (2005a). Popis morfologických značek – poziční systém. In: *Manuál korpusového manažeru Bonito*. [online], [cit. 2009-05-29]. Available from: <http://www.korpus.cz/bonito/znacky.php>

[7] Hajič, J. et al. (2005b). *A Manual for Morphological Annotation – Positional Tags*. [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/ch02s02s01.html>

[8] Hladká, B. (2000). *Czech Language Tagging*. PhD Thesis. Charles University, Prague.

[9] *MorphCon – Convertor of Czech Morphological Tagsets*. [online], [cit. 2009-05-29]. Available from: >http://morphcon.webnode.cz>

[10] *Natural Language Processing Centre* (Faculty of Informatics, Masaryk University, Brno). [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/en/nlplab>

[11] Osolsobě, K. (1996). *Algoritmický popis české formální morfologie a strojový slovník češtiny* (unpublished PhD Thesis). Brno.

[12] *PDT 2.0 Guide*. [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/index.html>

[13] Petkevič, V. *Popis morfologických značek použitých v korpusu orw-mte*. [online], [cit. 2009-05-29]. Available from: <http://ucnk.ff.cuni.cz/orwell_znacky.php>

[14] Rychlý, P. (2000). *Korpusové manažery a jejich efektivní implementace*. (PhD Thesis) Brno. [online], [cit. 2009-05-29]. Available from: <http://www.fi.muni.cz/~pary/dis.pdf>

[15] Sedláček, R. (1999). *Morfologický analyzátor češtiny*. (MA Thesis) Brno. [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/projekty/ajka/ajka.pdf>

[16] Sedláček, R. (2004). *Morfologický analyzátor češtiny Ajka.* [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/projekty/ajka/ajkacz.htm>

[17] Sedláček, R. (2006). *AJKA tagset*. Brno. [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/projekty/ajka/tags.pdf >

[18] *The Czech National Corpus*. [online], [cit. 2009-05-29]. Available from: <http://www.korpus.cz/english/index.php>

[19] *The Prague Dependency Treebank*. [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/pdt2.0/>

[20] *ÚFAL Tools* (Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics, Charles University, Prague). [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/tools.html>

[21] Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the Language Resources and Evaluation Conference, LREC 2008*. CD full edition + printed Conference Abstracts. Marrakech, Morocco. [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/~zeman/publikace/2008-02/tagdrivers-marrakech-styl-lrec.pdf>

[22] Zeman, D. *DZ Interset*. [online], [cit. 2009-05-29]. Available from: <https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset>

[23] Zeman, D. *DZ Interset – Features*. [online], [cit. 2009-05-29]. Available from: <https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset:features>

[24] Zeman, D. *DZ Interset – Tag Set Drivers*. [online], [cit. 2009-05-29]. Available from: <https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset:drivers>

Recent Developments in the National Corpus of Polish

Adam Przepiórkowski[1,5], Rafał L. Górski[2], Marek Łaziński[3,5], and Piotr Pęzik[4]

[1] Institute of Computer Science, Polish Academy of Sciences, Poland
[2] Institute of Polish Language, Polish Academy of Sciences, Poland
[3] Polish Scientific Publishers PWN, Poland
[4] University of Łódź, Poland
[5] University of Warsaw, Poland

Abstract. The aim of the paper is to present recent – as of July 2009 – developments in the construction of the National Corpus of Polish. The main developments are: 1) the design of text encoding XML schemata for various levels of linguistic information, 2) a new tool for manual annotation at various levels, 3) numerous improvements in search tools.

1 Introduction

The aim of the paper is to present recent – as of July 2009 – developments in the construction of the *National Corpus of Polish* (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; `http://nkjp.pl/`). The project, funded by the Polish Ministry of Science and Higher Education (project number: R17 003 03), was launched at the very end of 2007 and will run until the end of 2010.

The rest of this section presents the background of the project, the consortium, and envisaged applications of the results of the project in the humanities. The following section, §2, describes the expected results of the project in general terms.[6] Section 3 presents recent developments concerning corpus annotation and search tools, as well as text encoding standards deployed in the project. Section 4 concludes the paper.

1.1 Background

For Polish, the most represented Slavic language of the EU, there still does not exist a national corpus, i.e., a large, balanced and publicly available corpus, which would be at least morphosyntactically annotated. Currently, there exist three contemporary[7] Polish corpora which are – to various extents – publicly available. The largest and the only one that is fully morphosyntactically annotated is the IPI PAN Corpus (`http://korpus.pl/`; Przepiórkowski 2004), containing over 250 million segments (over 200 million

[6] These two sections overlap to a large extent with Przepiórkowski *et al.* 2008.

[7] Another – much smaller and dated, but historically very important – corpus is available in its entirety: `http://www.mimuw.edu.pl/polszczyzna/pl196x/`, a 0.5-million word corpus created in the 1960. as the empirical basis of a frequency dictionary (Kurcz *et al.* 1990).

orthographic words), but – as a whole – it is rather badly balanced.[8] Another corpus, which is considered to be carefully balanced, the PWN Corpus of Polish (`http://korpus.pwn.pl/`), contains over 100 million words, of which only 7,5 million sample is freely available for search. The third corpus, the PELCRA Corpus of Polish (`http://korpus.ia.uni.lodz.pl/`), also contains about 100 million words, all of which are publicly searchable.

1.2 Consortium

The project is unique in that it involves all major corpus developers for a given language, including the three developers of the three corpora mentioned above: the Institute of Computer Science at the Polish Academy of Sciences (ICS PAS; the Polish acronym is IPI PAN, hence the name of the corpus) in Warsaw, which coordinates the project, the PWN Publisher in Warsaw and the PELCRA group at the University of Łódź. The fourth partner is the Institute of Polish Language at the Polish Academy of Sciences (IPL PAS) in Cracow, the developer of an internal corpus, available only for research carried out at the Institute.

Another institution whose staff and students are involved in the project is the University of Warsaw. Some of the ideas which influenced the methodology of the project evolved in the Faculty of Polish Studies and in the Institute of Informatics, where two of the authors lecture on corpus linguistics, linguistic engineering and Polish grammar.

1.3 Practical applications in the humanities

The project is correlated with another national project, led by IPL PAS, aiming at the development of a new large dictionary of Polish, for which the National Corpus of Polish will serve as the empirical basis.

External users, such as lexicographers and linguists, will mainly be interested in searching the corpus for word and phrase concordances as well as collocations. The corpus can also serve as the treasure of well-known quotations from Polish and key words of Polish culture, with some emphasis on good representation of secondary school required readings in Polish literature and history. Therefore, quotations from the corpus will be crucial for new large dictionaries of Polish (including the new dictionary currently developed at IPL PAS, as well as dictionaries published by PWN), not only as a source of the typical uses of words, but also as a reference to cultural authorities rooted very well in the Polish literary tradition.

The corpus as a whole also enables creating a number of comparable corpora. The size of the corpus, exceeding the informal standard of 100 million, shall guarantee that there will be sufficient number of texts of different genres to meet the selection criteria of target corpora. Technically, there are two possible ways of creating a comparable corpus: to create a separate subcorpus comparable to a corpus of a different language or to place in the header of selected texts an element stating that this very text is a part of a comparable corpus. Hence, the user will have the option of narrowing down their query

[8] There exists a 30-million segment subcorpus of the IPI PAN Corpus which is relatively balanced.

to the texts marked as included in the comparable corpus, in the same way as they can narrow down the query to a certain author, genre or period.

A subproject of the NKJP consists in monitoring the words in daily and weekly newspapers and comparing word frequencies in two periods of time in order to test the saliency of words in every day public discourse. A demo version of this tool has recently been made available on the web site of the Association of Local Press (`http://www.gazetylokalne.pl/`). The project "Words of the Week" will be launched soon in cooperation with the weekly "Polityka" and later with Polish dailies. (A similar project was in progress on the website of the daily Rzeczpospolita from 2004 to 2007, see Łaziński and Szewczyk 2006).

2 Corpus

The intended size of the whole National Corpus of Polish is 1 billion words, of which at least 300-million word subcorpus will be carefully balanced.

2.1 Representativeness

In establishing the criteria of representativeness we build on our own experience, as well as on the experience of other national corpora, especially, the Czech National Corpus. We understand representativeness as representing the perception of language by a certain linguistic community (cf. Čermák *et al.* 1997), what in practice means reflecting the structure of readership in the structure of the entire corpus. There are theoretical reasons for this choice, besides practical considerations. Out of potential concepts of representativeness which may be applied to various corpora, the following two have the best methodological motivation: representing the population of texts or representing the structure of readership. If we adopted the model representing the production of text, around 90% of the corpus would consist of press. Hence the corpus would be representative but not balanced. The use of such a corpus in linguistic and lexicographic work is questionable (Górski 2008). The representativeness is based on several pools exploring the choices of reading of the average Polish reader as well as the circulation of the press. On the other hand we expect the corpus to be not only representative but also balanced, therefore the amount of press is lowered compared to the overall picture of the readership, so as not to let any text type dominate over the entire corpus (Górski 2009).

A first step towards establishing representativeness is determining the typology of text types. The typology is generally based on the bulk of work on Polish stylistics. There are however some text types which seem to be overseen by traditional stylistics, which had to be added. The classification is based mainly on intralingusitic features of texts. A small pilot study was conducted, so as to establish a set of purely linguistic factors differentiating the text types (Górski and Łaziński 2008). We are however aware of the fact, that the proposed classification is not the only possible.

As assigning a text to a certain type is not always a straightforward task, we decided that every text will be classified in a double-blind procedure by two linguists. In case of discrepancy they will have to discuss their decision.

To assure a wide coverage by topic we use classifications used by the libraries which will be encoded in the header of each text. To meet the needs of philologists and lexicographers we shall try as far as possible try to include in the corpus the most important works of modern Polish literature including poetry.

2.2 Spoken component

A 30 million word component of the NKJP will represent the spoken register of Polish. This part of the project is coordinated by the PELCRA team and it derives from the experience of compiling the 600 000 spoken-conversational component of the PELCRA corpus (Waliński and Pęzik 2007). Apart from transcripts of public speeches, parliamentary commission proceedings, televised debates, chat shows, radio interviews and news bulletins, a 3 million word subset of the spoken component will comprise natural, spontaneous conversations recorded by persons trained to preserve the natural character of the language data collected. Spoken NKJP data will be annotated with sociolinguistic metadata, including information on the age, gender, education and social background of the recorded speakers. Selected fragments of the spoken corpus will be aligned with the recordings and integrated in a relational database engine on top of which a publicly accessible web interface will be implemented (Pęzik *et al.* 2004).

2.3 Annotation

The entire corpus will be annotated linguistically, structurally and with bibliographic metadata. The basis of the linguistic annotation will be the full morphosyntactic annotation (not just parts of speech, but also values of cases, genders, etc., as appropriate). As in the IPI PAN Corpus, each segment (token) will contain not only the information about which morphosyntactic interpretation is correct in a given context, but also about all the other possible interpretations, rejected in the context (Przepiórkowski *et al.* 2004). Apart from the morphosyntactic annotation, the corpus will contain partial syntactic information, i.e., main types of syntactic groups will be identified, as well as named entities and various kinds of lexical constructions (so-called syntactic words). Also, forms of over 100 frequent semantically ambiguous lexemes will be disambiguated.

Because of the size of the corpus, it will not be possible to annotate the whole corpus manually. However, a 1-million word subcorpus of the representative 300-million subcorpus, reflecting its structure, is being annotated manually and it will be utilised for training and testing of linguistic tools which will subsequently be used for the automatic annotation of the whole corpus.

3 Recent developments

3.1 Text encoding

The need for text encoding standards for language resources (LRs) is widely acknowledged: within the International Standards Organization (ISO) Technical Committee 37 / Subcommittee 4 (TC 37 / SC 4; `http://www.tc37sc4.org/`) work in this area has

been going on since early 2000s, and working groups devoted to this issue have been set up in two current pan-European projects, CLARIN (`http://www.clarin.eu/`) and FLaReNet (`http://www.flarenet.eu/`). It is obvious that standards are necessary for the interoperability of tools and for the facilitation of data exchange between projects, but they are needed also within projects, especially where multiple partners and multiple levels of linguistic data are involved.

NKJP is committed to following current standards and best practices, but it turns out that the choice of text encoding for multiple layers of linguistic annotation is far from clear. Przepiórkowski and Bański 2009 contains an overview of recent and current standards and best practices, at various stages of development. The conclusion of this paper is that the guidelines of the Text Encoding Initiative (Burnard and Bauman 2008; `http://www.tei-c.org/`) should be followed, as it is a mature and carefully maintained *de facto* standard with a rich user base. Various proposed standards proposed by ISO TC 37 / SC 4, including Morphosyntactic Annotation Framework (MAF), Syntactic Annotation Framework (SynAF) and Linguistic Annotation Framework (LAF), are still under development and, especially SynAF and LAF, have the form of general models rather than specific off-the-shelf solutions.

Nevertheless, when selecting from the rich toolbox provided by TEI, an attempt has been made to follow recommendations of these proposed ISO standards, as well as other common XML formats, including TIGER-XML (Mengel and Lezius 2000) and PAULA (Dipper 2005). This work, described in more detail in Przepiórkowski and Bański 2009 (cf. also Bański and Przepiórkowski 2009), resulted in TEI P5 XML schemata encoding data models largely isomorphic with or at least mappable to those formats.

3.2 Annotation tools

As mentioned in §2.3 above, a 1-million word subcorpus of NKJP is being annotated manually and it will be used for training automatic annotation tools. Anotatornia, a tool for the manual annotation of word senses developed within a previous project at ICS PAS (Hajnicz *et al.* 2008), has been extended and extensively modified to allow for the manual addition of sentence boundaries, word-level segmentation, morphosyntactic annotation and word sense disambiguation (Przepiórkowski and Murzynowski 2009).

At each level, annotation is adduced by two linguists, connecting with the server via a web interface. In case of differences, both are notified that the other annotator made a different decision at a given place, but they are not informed about each other's decision. Each annotator may change their own annotation, or confirm it. If the discrepancy persists, a referee makes the final decision and suggests a modification of the annotation guidelines (Przepiórkowski 2009b), if necessary.

Currently annotation is performed at the levels of sentence segmentation, word-level segmentation and morphosyntax, with word sense disambiguation expected to start in August 2009.

For the manual morphosyntactic annotation, each segment is automatically marked with all interpretations known to the new version of Morfeusz (Woliński 2006), a morphosyntactic dictionary of Polish based on the data of the *Słownik gramatyczny języka polskiego* ('Grammatical dictionary of Polish'; Saloni *et al.* 2007). The task of the annotator is to select the right interpretation or add the correct interpretation, if it is not

among those proposed by Morfeusz. In the process, various deficiencies of Morfeusz and the underlying grammatical data base have been identified and corrected.

The morphosyntactic tagset used in NKJP is a modified version of the IPI PAN Tagset (Przepiórkowski and Woliński 2003a,b). The differences between the two tagsets are described – and a formal specification of the NKJP Tagset presented – in Przepiórkowski 2009a.

3.3 Search tools and NKJP demo

Since developing an efficient search tool able to manage a 1-billion corpus is a potentially high-risk task, two approaches are pursued in parallel.

The first approach is based on the combination of Apache Lucene (`http://lucene.apache.org/`) and relational database technologies, and it is partly inspired by the implementation of the PELCRA Corpus of Polish: we expect this approach to scale well with the size of the corpus. Apart from scalability, this search engine also focuses on providing convenient access to concordance and collocation search results in a variety of output formats, including downloadable spreadsheets, compressed URL-s, integrated browser plugins and web services. It has yet to be seen to what extent this approach will accommodate more complex types of linguistic search at various levels of annotation.

The second approach is based on Poliqarp (Janus and Przepiórkowski 2007a,b), a dedicated search engine developed at ICS PAS and currently serving a corpus of 250 million segments: while Poliqarp involves a very expressive query language, currently further expanded to accommodate syntactic queries, it is not clear how well it scales with the size of the corpus. So far, modifications of Poliqarp within NKJP consisted in developing a new corpus compiler, translating the TEI-based XML encoding of texts to an efficient binary format. End-user improvements include more specific error messages in case of not well formed queries, and an option to randomise search results.

Both search engines are successfully employed in the NKJP Demo (`http://nkjp.pl/index.php?page=6&lang=1`), which currently consists of about 500 million words.

4 Conclusion

As any "recent developments" publication, this paper describes work in progress. Intensive work within the National Corpus of Polish project concerns all levels of corpus development: from data acquisition, through text encoding and linguistic annotation, to efficient corpus search engines. We hope this overview paper has whetted the reader's appetite for the final project results.

References

Bański, P. and Przepiórkowski, A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009*, Singapore.

Burnard, L. and Bauman, S., editors (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford. `http://www.tei-c.org/Guidelines/P5/`.

Čermák, F., Králík, J., and Kučera, K. (1997). Recepce současné češtiny a reprezentativnost korpusu. *Slovo a slovesnost*, **58**, 117–124.

Dipper, S. (2005). Stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin.

Górski, R. L. (2008). Representativeness of the written component of a large reference corpus of Polish. Primary notes. Forthcoming.

Górski, R. L. (2009). The representativeness of NKJP. Talk delivered at *Practical Applications in Language and Computers (PALC 2009)*, Łódź, April 2009.

Górski, R. L. and Łaziński, M. (2008). Wzór stylu i wzór na styl. Zróżnicowanie stylistyczne tekstów Narodowego Korpusu Języka Polskiego. Talk delivered at the *VII Forum Kultury Słowa*, Gdańsk, October 2008.

Hajnicz, E., Murzynowski, G., and Woliński, M. (2008). ANOTATORNIA – lingwistyczna baza danych. In *Materiały V konferencji naukowej InfoBazy 2008, Systemy * Aplikacje * Usługi*, pages 168–173, Gdańsk. Centrum Informatyczne TASK, Politechnika Gdańska.

Janus, D. and Przepiórkowski, A. (2007a). Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In J. Waliński, K. Kredens, and S. Goźdź-Roszkowski, editors, *The proceedings of Practical Applications in Language and Computers PALC 2005*, Frankfurt am Main. Peter Lang.

Janus, D. and Przepiórkowski, A. (2007b). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 85–88, Prague.

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., and Woronczak, J. (1990). *Słownik frekwencyjny polszczyzny współczesnej*. Wydawnictwo Instytutu Języka Polskiego PAN, Cracow.

Łaziński, M. and Szewczyk, M. (2006). Słowa klucze w semantyce i statystyce. słowa tygodnia „Rzeczpospolitej". *Biuletyn Polskiego Towarzystwa Językoznawczego*, **LXII**, 57–68.

Mengel, A. and Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, pages 121–126, Athens. ELRA.

Pęzik, P., Levin, E., and Uzar, R. (2004). Developing relational databases for corpus linguistics. In B. Lewandowska-Tomaszczyk, editor, *The proceedings of Practical Applications in Language and Computers PALC 2003*, Frankfurt am Main. Peter Lang.

Przepiórkowski, A. (2004). *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Przepiórkowski, A. (2009a). A comparison of two morphosyntactic tagsets of Polish. In *Proceedings of the Mondilex workshop in Warsaw, June 2009*.

Przepiórkowski, A. (2009b). Zasady znakowania morfosyntaktycznego w NKJP. Version 1.19 of 27 July 2009.

Przepiórkowski, A. and Bański, P. (2009). Which XML standards for multilevel corpus annotation? Unpublished manuscript.

Przepiórkowski, A. and Murzynowski, G. (2009). Manual annotation of the National Corpus of Polish with Anotatornia. Talk delivered at *Practical Applications in Language and Computers (PALC 2009)*, Łódź, April 2009.

Przepiórkowski, A. and Woliński, M. (2003a). A flexemic tagset for Polish. In *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, pages 33–40, Budapest.

Przepiórkowski, A. and Woliński, M. (2003b). The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, pages 109–116.

Przepiórkowski, A., Krynicki, Z., Dębowski, Ł., Woliński, M., Janus, D., and Bański, P. (2004). A search tool for corpora with positional tagsets and ambiguities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 1235–1238, Lisbon. ELRA.

Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B., and Łaziński, M. (2008). Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech. ELRA.

Saloni, Z., Gruszczyński, W., Woliński, M., and Wołosz, R. (2007). *Słownik gramatyczny języka polskiego*. Wiedza Powszechna, Warsaw.

Waliński, J. and Pęzik, P. (2007). Web access interface to the PELCRA referential corpus of Polish. In J. Waliński, K. Kredens, and S. Goźdź-Roszkowski, editors, *The proceedings of Practical Applications in Language and Computers PALC 2005*, pages 65–86, Frankfurt am Main. Peter Lang.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 511–520. Springer-Verlag, Berlin.

Spoken Texts Representation in the Russian National Corpus: Spoken and Accentologic Sub-Corpora

Svetlana Savchuk

Vinogradov Institute of Russian Language of the Russian Academy of Sciences
Moscow, Russia
savsvetlana@mail.ru

Abstract. The paper presents the main characteristics of the Spoken and Accentologic corpora, available on open access on the site ruscorpora.ru. Corpora size, type and sources of text material, the way it is represented in the corpora, types of linguistic annotation, corpora composition and ways of their effective use according to their purposes are being described.
Keywords: spoken corpora, spoken data representation, linguistic annotation, accentology, the Russian National Corpus

1 Introduction

Modern corpora as a rule contain a spoken component, because it is oral speech that forms the core of linguistic performance. The interest in investigation of spoken data is constantly increasing in corpus linguistics, as referred to [1]. There are common recommendations and standards concerning the preparation and representation of oral texts in the corpus (TEI, EAGLES). Meanwhile each national corpus (British [2], Czech [3], Slovak [4], Russian [5], American [6] and so on) offers its own way of texts selection and corpus architecture.

As far as the Russian National Corpus is concerned, it has two distinctive features in organization of the spoken component. 1) Spoken corpus is excluded from the main body of written texts and forms a separate module in the RNC supplied with the types of annotation specific for spoken texts only. 2) We have refused the idea of creating a "universal" spoken corpus suitable for all kinds of oral speech study – from phonetic aspects to discourse analyses. That's why there is not one but three spoken sub-corpora within the RNC, with their own features and spheres of application. These are a) Spoken sub-corpus, b) Accentologic sub-corpus (both are functioning) and c) Multimedia Corpus (is under development).

The paper presents the main parameters of the Spoken and Accentologic corpora.

2 The Corpus of Spoken Russian

The Corpus of Spoken Russian is being created in cooperation with different research centers: Moscow, Saint Petersburg, Saratov, Perm, Yekaterinburg, Omsk, Krasnoyarsk, Uljanovsk, *etc*. Nowadays it accumulates both transcriptions of present-day spoken speech and data collected and described by the researchers since the 1960s.

The Corpus of spoken texts which functions as a part of the RNC provides researchers with ample opportunities.

2.1 Criteria for text selection

1) The Spoken corpus includes full authentic texts. It enables us to discover everything which could have been missed if we studied only selective records.

2) The corpus includes an amount of texts which significantly exceeds the average number of texts at the disposal of a colloquial speech researcher. It makes possible to explore frequency and randomness of a phenomenon, discover regular patterns which could only be found within a huge amount of texts and draw statistically reliable conclusions from these patterns.

3) The corpus contains texts heterogeneous in terms of when and where the records were made. They also differ in regard to sex- and age-related, social and professional features of speakers. It should be mentioned that dialectal texts are not included in the Spoken corpus, they are collected in the Dialectal corpus and supplied with inherent annotation.

4) The texts in the spoken corpus belong to a long time span of more than 70 years (if we take the scripts of the 1930s movies into account). The first records of spoken language date from 1956, the latest were made in spring 2009. It enables us to follow changes in colloquial speech (which are very swift), note any new trends, *etc*.

5) The corpus (in contrast to collections normally used for colloquial speech studies) includes spoken texts from different spheres of communication, which have been pronounced under different circumstances. We do not share some researchers' opinion that only 'natural speech of city residents in direct contact' may be considered as 'real Russian speech'[7]. Spoken speech as a form of a language's existence (as opposed to written form) is presented in different spheres of functioning: in everyday life sphere – as private (non-public) spoken speech; in educational and scientific sphere – as academic spoken speech; in journalism – as public spoken speech, TV speech and radio speech; in official sphere – as official spoken speech; in industrial and technical sphere – as professional spoken speech; in theological sphere – as a sermon; in advertising sphere – as TV and radio advertisements; in fiction – as speech of theatre and cinema. That is why a spoken text in the corpus is not only a dialogue in a shop or a family conversation at dinner table, but also an academic lecture, a report at a seminar, an author's meeting with readers, an interview or a TV talk-show, sports commentaries and some.

6) Another criterion for distinguishing different types of spoken speech, which is used for selecting corpus texts, is a level of preparation or spontaneity. We can use this scale to place different kinds of spoken texts in decreasing order in terms of spontaneity. The extreme points on this scale are spontaneous, non-prepared speech (dialogue and monologue) on the one hand and written-to-be-spoken speech (reading a text aloud, pronouncing a text learned by heart) on the other hand. There are a lot of text types between these points which were called quasi-spontaneous [8]: interview, narrative on

foreknown topic, speech prepared according to premeditated plan, retelling of another person's speech, formulaic speech, *etc*. In the corpus there are no spoken texts which would represent written-to-be-spoken speech. But there are a lot of texts which would be estimated as quasi-spontaneous on this scale. First of all, these are the records of public speech and the movie sub-corpus.

7) The movie sub-corpus includes scripts of live-action movies and animated films (and in future also documentaries and commercials). It is a unique component of the Spoken sub-corpus in the RNC. This language sphere until recently has been escaping spoken language researchers' and corpora authors' attention. Meanwhile, the influence of these texts on Russian (and not only Russian) language usage is great, as it has been shown in [9].

2.2 Composition and structure of the Spoken corpus

At present the spoken corpus consists of 8.6 mln tokens. It may be considered as a representative text collection, which reflects the functioning of contemporary Russian spoken language. Let us show what effect it has on the corpus composition and structure[1]. The distribution of spoken texts according to main parameters is shown in tables below.

Sphere	Tokens	Percentage
Public speech	4392123	51
Private speech	780150	9
Speech of cinema	3378096	40

Table 1. The distribution of texts according to spheres of spoken communication

Sphere	Text types	Tokens	Percentage
Public speech	discussion	1920306	43.7
	talk	1232823	28.1
	interview	309918	7.1
	lecture	195302	4.4
	sports commentary	182811	4.2
	narration	112820	2.6
	parliamentary hearings	86640	2
	conference	48972	1.1
	round table	49177	1.1
	other	253354	5.7
Private speech	conversation	609167	78.1
	telephone conversation	80648	10.3
	tale	47340	6.1
	retelling	12533	1.6
	microdialogues	25435	3.3
	argument	5027	0.6

Table 2. The distribution of spoken texts according to text types

1 Data are given as of July 2009.

Sphere	Text types	Tokens	Percentage
Speech of cinema	drama	844713	25.0
	comedy	1194060	35.3
	detective	313611	9.3
	films for children	258368	7.6
	costume films	112487	3.3
	science fiction and fantasy	93156	2.8
	action	57823	1.7
	animated films	102882	3
	other	400996	12.0

Table 3. The distribution of speech in movies according to film genres

The corpus includes texts of different *topics*. The texts with the mark 'private life' are the most frequent (more than 50% of all texts). Politics and social life texts, art and culture texts, scientific texts, leisure and entertainment texts, sport texts are less frequent in decreasing order.

As for the time of recording, the majority of texts dates from 2003 to 2008, the other significant part of the texts (more than 600 thousand tokens) are from the 1990s, 947 thousand from 1980s, about 1,9 mln from 1960-1970s, more than 420 thousand – before 1960 (the last are mainly movie transcripts) .

Period	Percentage
1930–1959	5
1960–1979	22
1980–1989	10
1990–1999	8
2000–2008	55

Table 4. The distribution of spoken texts according to date of recording

Geographical distribution of the spoken corpus is quite broad. The corpus includes texts recorded in Moscow and Moscow region (the majority of texts), in Saint Petersburg, Kirov region, Saratov, Samara, Taganrog, Voronezh, Novosibirsk, Uljanovsk, Yekaterinburg.

2.3 Language data presentation method

Texts for the Spoken corpus are derived from various sources. A considerable part of them are records of spoken texts published in reading books and collections compiled by the researchers of spoken language and edited by E. A. Zemskaja, O. A. Lapteva, N. N. Rozanova and M. V. Kitaigorodskaja, A. S. Gerd, O. B. Sirotinina and others in the 1970-1990s. Another part are records of spoken texts collected in such research centers as Vinogradov Institute of Russian language, Moscow State University, Saint Petersburg State University, Saratov University, Uljanovsk University. Transcripts of sociologists' conversations in focus groups, based on different important social topics

form an important part of public spoken speech in the Corpus. These records were rendered by the Public Opinion Foundation. All movie transcripts and a high percentage of records of spoken texts are made by the members of the RNC team or under their supervision.

All texts are represented in modern orthography. The syntagm boundaries are marked by slashes. Punctuation marks (dot, elision, interrogation and exclamation marks) are used for marking the end of a phrase. Some typical features of oral speech are reflected in transcripts. This involves an increase in word-forms with non-standard spelling. The issue is dealt with preliminary *orthographical normalization* of transcripts. All cases of substandard spelling are supplied with their standard equivalents which are recognized and annotated later with grammar features attributed to the whole complex. As a result on screen display we can see the peculiarities of pronunciation, captured by non-standard word-forms (*поскоку, тыща* in the example below) with the correct grammar parsing (because the normalized forms *поскольку*, *тысяча* are analyzed). You would get the same if you input these words in the corpus dictionary and had them identified correctly by the lemmatizer.

> [Васильев, муж, 1962] но вот смотрим / Аня Бога́лий / она в общем-то недалеко / она рядышком / как я уже говорил / основна… основные события будут развиваться на огневом рубеже / *поскоку* / здесь уже я повторяюсь / но для тех / кто не слышал / здесь в Сан-Ска́рио очень сложная высота / практически почти *тыща* семьсот метров над уровнем моря / и дышится ну крайне тяжело / очень сильно разря́женный воздух *[Константин Выборнов, Дмитрий Васильев. Спортивный репортаж: биатлон. Олимпийские игры. Эстафета 4х6 км Женщины (23.02.2006) // «Первый канал», прямой эфир, 2006].*

2.4 Types of linguistic annotation

The same types of annotations as in the RNC are typical of the Spoken corpus – metatextual, morphological and semantic. Thereby, the same types of subcorpora and the same types of search as in 'the written corpus' are available in the spoken corpus.

Metatextual annotation marks a text as a whole and includes information regarding author's name, sex, age or date of birth, text characteristics (creation date, sphere of functioning, text type, movie genre, domain) and so on [10].

Morphological information is assigned to a word-form and consists of four groups of tags: 1) lexeme (a dictionary form of the lexeme and the part of speech to which it belongs); 2) a variety of the lexeme's grammatical features, known as word-classifying features; 3) a variety of the word-form's grammatical features, known as word-altering features; 4) information concerning non-standard forms of the word-form, orthographic variations, *etc*.

Sociological annotation is specific to the spoken corpora only. It is assigned to different speaker's utterances and characterizes a word usage from the point of view of sex and age of a speaker (if this information is available). Sociological annotation

allows a user to create his/her own sub-corpora by various parameters or their combinations: by a speaker's sex (so a user could create a sub-corpus of feminine or masculine spoken language); by a speaker's age (for example, a user can create a sub-corpus of teenagers' phrases); by a speaker's year of birth (this option is available only for movie transcripts, so you could select the phrases by the actors born in 19th century); by an actor's name (for example, you can create a sub-corpus of Eugene Leonov's phrases).

Apparently, sociological annotation may be supplemented with metatextual annotation which makes it possible to select texts by one speaker and include his/her name and year of birth in the description of the text. It is clear, that if a) there are more than one speaker, b) speakers cannot be named because of ethical reasons, c) their age is unknown or speakers are of very different age, this information cannot be included in the description of the text. In this case all we can refer to is sociological annotation.

2.5 Prospects of development

As it has been mentioned above, spoken language in the RNC is presented in three corpora. In terms of data representation the spoken corpus is apparently at a disadvantage comparing to both the accentologic corpus (because it has information about stress) and the multimedia corpus (because it portraits real language). But still, does it have any prospects of development? We should answer 'yes' to this question because of several reasons.

First of all, the spoken corpus differs from the accentologic and multimedia corpora in text types. As we have already mentioned, there are spoken texts recorded in different regions of Russia at different times. In fact, if we had records, money and staff nothing could deter us from matching transcripts with audios and process the texts the same way as in the accentologic corpus.

However, it is not always possible. The significant part of spoken texts (first of all, earlier records and collections from regional research centers given to the RNC) exist only as transcripts. Tape records either have been lost or have never been made (if microdialogues were recorded manually). In the first place it applies to the texts published in reading books [11], [12], [13]. These records can only be presented as a part of the spoken corpus.

In spite of the improvement of equipment making transcripts of audio records is still the most widespread and the most reliable way to put down spoken texts. This resource shouldn't be underestimated. As the experience of practical training in collecting spoken texts by Moscow students has shown, transcripts are not always supplemented with satisfying audio records which could be used in the corpus. There are different reasons for this – technical or just accidental: poor quality of the records, rare file formats, conversion errors, *etc*. Such records cannot be used in the accentologic corpus, but they can take place in the spoken corpus. Thereby, the

spoken corpus excels both the accentologic and upcoming multimedia corpora in the number and diversity of texts.

The second reason for continuing development of the spoken corpus is the nature of linguistic annotation and the corpus search engine. The corpus is an effective research tool only if its annotation corresponds to the linguistic goals defined by a researcher. Thus, researches of most morphological, syntactic and lexical features of spoken language are easier to conduct using the materials of the spoken corpus. Among its main advantages are its size, diversity of texts, annotation similar to the one in the written corpus, which allows to compare results easily. When researching phonetic, accentologic, prosodic and paralinguistic features of spoken utterances, it is better to refer to the accentologic or multimedia corpora.

Thus, the most urgent problem of the spoken corpus development is increasing of its size up to 10 million tokens. This can be achieved by balancing the corpus and adding texts which are not yet very well presented – private speech in the first place. Another goal is to broaden the geography of the corpus by adding Russian spoken texts from different regions, the near abroad and foreign countries. This would enable us to study the Russian language which functions in contact with related, non-related and foreign languages.

3 The Accentologic corpus of Russian

Accentologic corpus gives an opportunity to study word stress. This information is very important for languages with non-fixed stress. The Russian language is one of them. It has a very complicated accentual system. Russian stress has the following features: firstly, it is non-fixed, which means that any syllable may be stressed (for example, *зо'лото, воро'на, борода'*); secondly, it is variable, which means that it may change its place as a result of inflexion or word formation (for example, *зо'лото – золото'й*, *золоти'ть*, *позоло'та*; *борода'* (n, f, nom, sg) – *бороды'* (n, f, gen, sg), *бо'роду* (n, f, acc, sg), *бо'роды* (n, f, nom, pl), *боро'дка* (diminutive form).

Moreover, the Russian stress system is in process of rearrangement, when significant changes are rather swift. That is why from the very beginning the Accentologic corpus was planned as diachronic, which would allow studying the history of the Russian stress. Meanwhile, there are no limitations of the size of the corpus (except technical capabilities). The goals of the corpus define its structure, composition and criteria for text selection.

3.1 Criteria for text selection

There are two zones in the Accentologic corpus.

1) The zone of *prose* includes oral texts and films transcripts, in which stressed syllables are marked according to real pronunciation. The main criterion for including a text in the corpus is the availability of a record of sufficient quality which would allow us to verify the transcript. We are equally interested in both accentologic

standards of the literary language and their variants which change over a period of time, because text annotation makes it possible to correlate any variant to a sphere of functioning, a genre and a speaker. Thus, the prose zone contains some examples of spontaneous everyday speech, spoken public speech of different levels of spontaneity (TV and radio speech, political speeches, academic spoken speech, sermons, *etc*.), movie and radio play transcripts, reading aloud. The earliest records of this zone date from the beginning and the first decades of the 20th century (gramophone records of L.N. Tolstoy's letters, political leaders' speeches, records of speeches made at the First congress of writers in 1934, films of 1930s). In perspective some of accented written texts (for example, books and manuscripts of the 18th and 19th centuries, and later – even older texts) may be included in this zone.

2) The zone of *poetry* contains texts with marked accented syllables, so one can define the exact word stress using special rules. Specially annotated poetic texts of the 18th-20th centuries are included in this zone and still continue to be added. At present this zone mainly reflects the history of the Russian stress, because the corpus contains poetry written before the 20th century.

3.2 Structure and composition of the corpus

Nowadays the Accentologic corpus contains more than 6.4 million tokens. Texts' distribution among the two zones and according to time periods is listed in the tables below.

Zone		**Tokens**	**Percentage**
Poetry		3238956	50.5
Prose	Speech of cinema	3198998	49.7
	Public speech	not available	
	Private speech	not available	
	Reading aloud	not available	

Table 5. Texts proportion in the zones of the Accentologic corpus

Zone	**Poetry**	**Prose**
1700 -1799	738148	
1800-1910	2500808	
1911-1949	-	248237
1950-1979	-	1537243
1980-1999	-	865503
2000-2008	-	548015

Table 6. Distribution of texts according to creation date

3.3 Types of annotation

The Accentologic corpus is supplied with four types of annotation which are used in the Spoken corpus, and besides that has its own accentologic mark-up. In *metatextual* annotation not only characteristics common to both zones of prose and poetry are

used (the author's name, age, sex, date of text recording/ creating), but also parameters intrinsic to each zone severally. These are genre, meter, clausula, rhyme, line size for poetry and text type for prose.

Each word in Accentologic Corpus has *morphological* and *semantic* annotation as in the RNC in whole. *Sociological* annotation which attributes the information about a speaker's sex, age (or date of birth) to each utterance is used in the zone of prose only.

Due to *accentologic* annotation each word is supplied with stress marks, so you can make different kinds of search requests and retrieve data of stressed or unstressed word-forms in combination with grammatical and semantic features.

3.4 Language data preparation and presentation method

Preparation of texts for Accentologic corpus is performed in several steps.
The first step includes audio files decoding, orthographical normalization and transcripts editing. The second stage is automatic and uses the software Accentuator (made by A. Polyakov) which accent words in accordance with the recommendations of the built-in lexicon. This lexicon is compiled from the database of normative Russian dictionaries, but is also amplified by the corpus developers. At the third stage an expert listens to audio records, reviews them against transcripts and corrects annotated texts. As a result we get a text accented the same way it is pronounced.

[Сергей Ненашев, Александр Абдулов, муж, 38, 1953] Ро́вно в четы́ре де́сять вы́ключи́те телеви́зор / настро́ив его́ на четвёртый кана́л. *[Виктор Сергеев, Игорь Агеев. Гений, к/ф (1991)*]

As we have mentioned before, in the zone of poetry the same annotation as in the Poetry corpus is used. The special programme annotates up beats and then an expert corrects any mistakes according to the special manual. As a result we get a text which looks like this:

Он идѐт в воротà, он ужѐ на крыльцѐ, Он взошѐл по круты̀м *ступеня̀м* На площà̀дку и вѝдит: с печà̀лью в лицѐ Одино̀ко-уны̀лая тà̀м *[В.А. Жуковский. Иванов вечер (1822)]*

This example is trisyllabic anapaest in which the marked up beats (ictuses) match the real word stress.

Вот на̀ш геро̀й подъѐхал к сѐням; Швейца̀ра мѝмо о̀н стрело̀й Взлетѐл по мра̀морны̀м *ступѐням*, Распра̀вил во̀лоса̀ руко̀й, Вошѐл. *[А. С. Пушкин. Евгений Онегин / Глава первая (1823-1824)]*

In this iambic tetrameter the number of up beats (ictuses) is bigger than the number of word stresses. In this case we should exclude all *impossible* stresses (which are not presented in any dictionaries or reference books) and take into consideration *possible* stresses only. For the word-form *мра'морным* the only possible stress is the stress on the first syllable, for the word-form *волоса'* (n, m, nom, pl) – the stress on the ending, but the stress on the first syllable would characterize this word-form as gen. sg. Interestingly enough, in these two examples the stresses of the form dat. pl. of the noun *ступеням* are different, which shows the coexistence of these variants in the beginning of the 19th century.

3.5 Intended use and prospects of development

The Accentologic corpus is one of specialized corpora and aims for researches in a quite narrow sphere of the Russian accentology. In spite of rather small size of the corpus it is still enough to give a researcher an opportunity to study new material, correct some points of his/her theory and normative recommendations [14].

Furthermore, the corpus enables us to solve methodological problems: it can be used as a reference material, as a material for compiling exercised and teaching aids for those learning Russian.

In the nearest future the increase of corpus size and text variety is planned. According to this plan poems of the 1st half of the 20th century will be included in the zone of poetry. The zone of prose will be expanded due to texts belonging to different spheres of spoken communication and created in various time periods. Among them are academic lectures, interviews and TV talk-shows, sports commentaries, sermons, political speeches, narratives, private conversations and so on. Many of these texts will be available on-line next year.

References

[1] Rayson, P., Mariani, J. Visualizing corpus linguistics. In: Corpus Linguistics 2009. 20-23 July 2009. Abstracts, p. 201. Liverpool (2009).
[2] British National Corpus, http://www.natcorp.ox.ac.uk/corpus/creating.xml.ID=design#plan
[3] Czech National Corpus, http://ucnk.ff.cuni.cz/english/struktura.php
[4] Slovak National Corpus, http://korpus.juls.savba.sk/structure/index.en.html
[5] Russian National Corpus, http://ruscorpora.ru/search-spoken.html
[6] American National Corpus, http://www.americannationalcorpus.org/ SecondRelease/ contents.html
[7] Matveeva, T. V. (ed). Zhivaja rech uralskogo goroda. Teksty (Живая речь уральского города. Тексты). Yekaterinburg Univ. Press, Yekaterinburg (1995)
[8] Galyashina, E.A. The differentiation features of prepared and spontaneous discourse in oral and written speech (Проблема дифференциации спонтанной и подготовленной речи) In: Papers from the Annual International Conference “Dialogue 2002”, http://www.dialog-21.ru/materials/ archive.asp?id=7287 &y=2002&vol=6077
[9] Grishina, E. A. Dva novych projekta dl’a Natsional’nogo korpusa: multimedijnyj podkorpus i podkorpus nazvanij (Два новых проекта для Национального корпуса: мультимедийный подкорпус и подкорпус названий). In: Natsionalnyj korpus pusskogo jazyka: 2003–2005. Rezul’taty i perspektivy. Indrik, Moscow (2005)

[10] Grishina, E. A. Ustnaja rech v Natsional'nom korpuse russkogo jazyka (Устная речь в Национальном корпусе русского языка) In: Natsional'nyj korpus pusskogo jazyka: 2003–2005. Rezul'taty i perspektivy. Indrik, Moscow (2005)
[11] Zemskaya, E. A., Kapanadze, L. A. (eds.) Russkaja razgovornaja rech: Teksty (Русская разговорная речь: Тексты). Nauka, Moscow (1978)
[12] Kitajgorodskaya, M. V., Rozanova, N. N. Rech moskvichej: Kommunikativno-kulturologicheskij aspect. (Речь москвичей: Коммуникативно-культурологический аспект). 2nd ed. Nauchnyj mir, Moscow (2005)
[13] Sergieva, N. S., Gerd, A. S. (eds.) Russkaja razgovornaja rech jevropejskogo severo-vostoka Rossii (Русская разговорная речь европейского северо-востока России). Syktyvkar (1998)
[14] Grishina, E. A. Natsional'nyj korpus russkogo jazyka kak istochnik svedenij ob ustnoj rechi. In: Rechevyje tekhnologii, 3. (2009)

The Meaning of the Conditional Mood Within the Tectogrammatical Annotation of Prague Dependency Treebank 2.0

Magda Ševčíková

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague, Czech Republic

Abstract. The conditional form is one of the moods of Czech verbs, and it renders several meanings in contemporary Czech texts (Sect. 2). The present paper focuses on the primary function of this mood, which is to express hypothetical events (Sect. 3). In Section 4, we briefly mention how modality has been treated up to now in PDT 2.0 and some other treebanks and finally in Section 5 we propose a new way how the primary meaning of the conditional mood should be captured in the annotation scheme of the tectogrammatical layer of PDT 2.0.

1 Introduction

Verbal mood as a universal morphological category occurs in many languages of the world, however, it is structured differently in different languages (see, e.g., [1], [10]). In Czech, three moods are traditionally distinguished: indicative, imperative, and conditional. In this paper, attention is focused on the Czech conditional mood, in particular on its primary function. As the corpus data document, the conditional is primarily used to refer to events which may be generally characterized as hypothetical; more specifically, to events which are contingent on realization of other events, to events which cannot happen any more etc. We want to demonstrate that in contemporary Czech the conditional as a means expressing hypotheticality of an event is a semantically relevant means which contributes to the meaning of the whole sentence. Thus, it has to be included in the representation of the meaning of the sentence.

In Prague Dependency Treebank 2.0 (PDT 2.0 in the sequel), the linguistic meaning of the sentence is represented as a tectogrammatical tree, which consists of nodes and edges with a set of attributes (see [6]). However, neither in PDT 2.0 nor within Functional Generative Description (FGD in the sequel; c.f. [14]), on the basis of which the annotation scenario of PDT 2.0 was built, any considerable attention has been paid to the semantics of the morphological category of mood.

After an overview of the functions of the conditional mood (Section 2), the semantic relevance of the conditional in its primary function is documented in Section 3 of the present paper. In Section 4, the treatment of the conditional and other modal means in PDT 2.0 (and in FGD) as well as in some other treebanks is described. A new formal means (called grammateme in PDT 2.0 and FGD) for capturing the primary function of the conditional within the tectogrammatical annotation of PDT 2.0 is proposed in Section 5. The paper concludes with some final remarks in Section 6.

2 Functions of the conditional mood

2.1 The primary function of the conditional

The morphological category of verbal mood is acknowledged to be one of the means by which modality is expressed in Czech. Modal meanings are further expressed by many other means such as modal verbs, modal particles, prosody etc. The primary function of the conditional is described under several terms in Czech linguistic literature. In this paper, we use the term hypotheticality to refer to the semantics of this mood. In its primary function, the conditional is opposed to the indicative by means of which events are simply asserted, presented as given. In Grammar of Czech (see [8], [2]), the semantic opposition of hypotheticality vs. assertion is called factual modality. The conditional mood is considered to be the marked member in this opposition.

Two types of hypothetical events are further distinguished: events which could happen (potential events) vs. events which cannot happen (irreal events).[1] Potential events are expressed by the so called present conditional as one of two subcategories of the conditional mood. The second subcategory of the conditional, the so called past conditional, expresses irreal events unambiguously. However, the present conditional is frequently (or even predominantly) used instead of the past conditional in Czech in the last decades. Such a substitution has been regarded as acceptable if the irreality is signalized by an adverb etc. (e.g. in [5]). Nevertheless, in the analyzed corpus data, the present conditional was used as an expression of irreal events also in cases in which the irreality could be resolved only on the basis of a very broad context or of knowledge of situation. Such an ambiguity causes substantial problems with annotation.

2.2 Secondary functions of the conditional

Besides the primary function, the conditional fulfills also other (secondary) functions in Czech. For example, in the sentences *Doporučil bych vám tu smlouvu podepsat* 'I would recommend you to sign the contract' or *Uvedl byste konkrétní příklad?* 'Would you give a concrete example?', the speaker uses the conditional to express his recommendation or appeal in a more polite way than using the indicative.

The conditional can further express the speaker's will to do something (c.f. *Kouřil bych* 'I would smoke', which actually means "I want to smoke") or his conviction whether the content will be realized or not (e.g., *Šel bych* 'I would go', which can be used instead of "I will probably go" in an appropriate context). This paper concerns with the primary function of the conditional; contexts in which it is used in its secondary functions are not analyzed here.

[1] The term 'irreal' is used according to [17], i.e., it refers to events which are impossible rather than to unreal events.

3 The conditional mood as a means for expressing hypothetical events

3.1 Corpus material analyzed

The primary function of the conditional mood was studied on the basis of language data from two corpora: from PDT 2.0 (`http://ufal.mff.cuni.cz/pdt2.0`; see [6]) and from SYN2005 corpus (`http://ucnk.ff.cuni.cz`).

PDT 2.0 is a collection of contemporary Czech newspaper texts to which a morphological annotation and annotation at two syntactic layers was assigned, at the so called analytical layer (layer reflecting the surface shape of the sentence) and at the tectogrammatical layer (layer of the linguistic meaning of the sentence). Annotation of all three types is available for more than 49 thousand sentences (i.e., more than 830 thousand tokens); the train data set (about 80 % of the full data) was used to search for sentences to be examined.

SYN2005 is a representative corpus of contemporary written Czech which contains 100 million tokens. In comparison with PDT 2.0, only morphological annotation was assigned to the data of SYN2005.

3.2 Semantic relevance of the conditional

In order to show the semantic relevance of the conditional mood in its primary function, we proposed a substitution test. In the test, the conditional, which is supposed to be the marked member of the opposition of the factual modality, was substituted for the indicative as for the unmarked member of this opposition. The substitution test was performed in several types of sentences from PDT 2.0 and SYN2005 corpus in the following way: in a concrete sentence which involved a present or past conditional verb form, the conditional was replaced by the indicative. Then, we tested whether the resulting sentence with the indicative can be used as an expression of the hypothetical event in question and whether it is acceptable with regard to the immediate as well as broader context, to our knowledge of situation etc. (c.f. [16] for more details on the substitution test).[2]

The acceptability of the indicative in the examined sentences seems to be related to the type of hypotheticality expressed by the sentence, i.e., whether a potential or an irreal event was expressed. Irreal events can be usually expressed only by the past or present conditional. For instance, in sentences (1) and (2), which both render irreal events, the present and past conditional verb forms are used, respectively; the conditional cannot be substituted for indicative verb forms here due to the given contexts, cf. in the sentence (1') the indicative is unacceptable due to the fact that Joseph Roth is not alive any more, in (2') since they (the photos) were not reminded.

On the contrary, when expressing potentiality, the present conditional can be often replaced by the indicative; see examples (3) and (3'), (4) and (4'). This substitutability is connected, in our opinion, with the semantic closeness of the potentiality of the conditional and the future meaning of the indicative (cf. [17]).

[2] In [7], the semantics of the past conditional was tested in a similar way.

(1) *Svatý pijan Joseph Roth* ***by*** *dnes* ***oslavil*** *rovnou stovku.* 'The saint drunkard Joseph Roth **would celebrate** his 100th birthday today.' (PDT 2.0)
(1') #*Svatý pijan Joseph Roth dnes* ***oslaví*** *rovnou stovku.* '#The saint drunkard Joseph Roth **celebrates** his 100th birthday today.'[3]

(2) ... *připomínat si je* ***by bylo bývalo*** *bolestivé.* '... to remind oneself of them **would have been** painful.' (SYN2005)
(2') #... *připomínat si je* ***bylo*** *bolestivé.* '#... to remind oneself of them **was** painful.'

(3) *Uhrát tu remízu* ***by bylo*** *úspěchem.* 'To draw the game **would be** a success.' (PDT 2.0)
(3') *Uhrát tu remízu* ***bude*** *úspěchem.* 'To draw the game **will be** a success.'

(4) *Podle A. Röschové (ODS)* ***by*** *pro Kozlův návrh* ***byla*** *asi polovina klubu ODS, ale osobně* ***by byla*** *spíš proti.* 'According to A. Röschová (ODS), approximately a half of the club of ODS **would vote** for Kozel's proposal but she personally **would vote** against it.' (PDT 2.0).
(4') *Podle A. Röschové (ODS)* ***bude*** *pro Kozlův návrh asi polovina klubu ODS, ale osobně* ***bude*** *spíš proti.* 'According to A. Röschová (ODS), approximately a half of the club of ODS **will vote** for Kozel's proposal but she personally **will vote** against it.'

3.3 Conditional vs. indicative in selected contexts

However, in the corpus data, several sentences expressing potential as well as irreal events occurred which violated our hypothesis of the relation of the hypotheticality type and the acceptability of the indicative. On the one hand, the indicative could be used instead of the conditional in some irreal contexts. On the other hand, the indicative was not acceptable in several sentences which expressed potential events. These two cases could be characterized as follows:

- An irreal event could be expressed both by the (present or past) conditional and the indicative, for instance, in governing clauses of the examined conditional complex sentences. In our opinion, the indicative comes into consideration in these sentences since the irreality is signalized explicitly by another means besides the verbal mood, e.g., by the conditional clause; see the sentences (5) and (5').
- Concerning the indicative in potential contexts, the loss of the hypotheticality which is connected with the examined substitution plays an indispensable role when judging the acceptability of the indicative instead of the conditional. The conditional mood cannot be replaced by the indicative in such sentences which, according to our interpretation, express potential events whose realization the speaker does not accept (cf. the example (6) vs. the sentence (6'), which cannot be accepted in the given context).

[3] The hash mark # indicates that the given sentence (with the indicative) in comparison to the original sentence (with the conditional) is unacceptable since it does not express a hypothetical event without regard to the fact that it is mostly a grammatical Czech sentence.

(5) *... kdyby dnes přišel Thomas Alva Edison do české banky se žádostí o úvěr na výrobu jakýchsi "žárovek, které změní svět", **skončil by** s nepořízenou ...* '... if Thomas Alva Edison would come to a Czech bank with an appeal for a credit for production of a kind of "light bulbs which change the world" in these days, he **would go away** empty-handed ...' (PDT 2.0)
(5') *... kdyby dnes přišel Thomas Alva Edison do české banky ..., **skončí** s nepoří-zenou ...* '... if Thomas Alva Edison would come to a Czech bank ... in these days, he **goes away** empty-handed ...'

(6) *V každém případě **by** v Polsku **vypukla** vážná politická krize.* 'In all cases a serious political crisis **would break** out in Poland.' (PDT 2.0)
(6') #*V každém případě v Polsku **vypukne** vážná politická krize.* '#In all cases a serious political crisis **breaks out** in Poland.'

To resume the results of the substitution test, the conditional cannot be replaced by the indicative in most sentences expressing irreal events and in such sentences rendering potential events whose realization is not expected by the speaker. These cases are considered to be an evidence that the conditional is an irreplaceable means in contemporary Czech texts and has to be included in the representation of the meaning of the sentence. On the contrary, the conditional can be substituted for the indicative in many sentences expressing potential events but also in some sentences with irreal meaning. The relatively free substitutability of the two moods in potential contexts is possible due to the close relation between the meanings of potentiality and future. In sentences which express irreal events, the conditional can be substituted for the indicative only if the irreality is clearly expressed by some other means; these cases do not deny the semantic relevance of the conditional.

4 Modality in PDT 2.0 and some other treebanks

4.1 Description of modality in PDT 2.0 and FGD

In theoretical works based on FGD, neither verbal mood nor other modal means have been studied yet in more detail in spite of the importance of these means in constituting sentence semantics. In PDT 2.0, a special grammateme verbmod was defined. However, since the values of this grammateme directly correspond to the morphological moods occurring in the sentence, the grammateme is not to be considered as a semantic counterpart of the morphological category of mood but rather as a provisional, "technical" solution which requires further investigation.

Within the broad area of modality, main attention has been paid to modal verbs in FGD (see esp. [12]). On the basis of syntactic and semantic criteria, a group of so called proper modal verbs were identified; e.g. *muset* 'must / have to', *moct* 'can'. Meanings of these verbs are considered as modal features "added" to the meaning of an autosemantic verb. In the tectogrammatical tree, meanings of proper modal verbs are represented by a grammateme belonging to the node representing the autosemantic verb in question (i.e., similarly to meanings expressed by the morphological category of tense etc.). Within PDT 2.0, the modality expressed by modal verbs was captured in the same way as in FGD.

4.2 Modality in some other treebanks

Also other treebanks, if paying any attention to modality at all, confined to modal verbs whereas semantics of verbal moods is omitted (the mood is usually reflected just within the morphological annotation). Semantics of modal verbs is described, e.g., in Proposition Bank, which is a corpus annotated with semantic roles, or in the annotated trebank of Bulgarian texts BulTreeBank. In Proposition Bank, a special semantic role MOD for modal verbs was included (cf. [11]). In BulTreeBank, deontic and epistemic reading of modal verbs was distinguished (see [9]). Also [3], [13], [4] or [15] were concerned with (semi-)automatic detection of deontic and epistemic usage of modal verbs, especially in scientific articles.

5 Representing the primary function of the conditional mood at the tectogrammatical layer of PDT 2.0

5.1 Grammateme of factual modality

As sketched in Section 3, the conditional in its primary function is a semantically relevant means: by the use of the conditional an event is presented as hypothetical. In this function, the conditional is opposed to the indicative, by means of which events are presented as given. The semantic opposition which is expressed by the conditional and the indicative is thus to be captured when representing the meaning of the sentence, i.e., within the tectogrammatical tree in PDT 2.0.

Since verbal mood is a morphological category (similarly to that of tense etc.), we propose to capture the semantic opposition expressed by the conditional and the indicative by a grammateme which belongs to a verbal node of the given tectogrammatical tree. We introduce a new grammateme of factual modality (factmod). Besides the opposition of hypotheticality and assertion also the difference between two types of hypothetical meanings – the potential and the irreal ones – has to be taken into consideration. Three values of the grammateme factmod are therefore proposed: potential for potential events, irreal for irreal events and asserted for given events.[4]

5.2 Annotation rules for assigning the grammateme of factual modality

When describing the assignment of the values of the proposed grammateme, we proceed from the surface structure of the sentence in a way similar to that in which a tectogrammatical representation is assigned to a sentence. A tectogrammatical node which represents an indicative verb form (the primary function of the indicative and the conditional is taken into consideration here and in the sequel) is typically assigned with the value asserted of the grammateme factmod. When a verb form of past conditional occurs, the value irreal is filled in the grammateme factmod at the tectogrammatical node representing this verb form.

[4] The solution to represent both oppositions by a single grammateme was preferred to the possibility to introduce two grammatemes just for the economy of the former solution.

verbal mood	grammateme factmod	grammateme tense
past conditional	irreal	–
present conditional expressing potentiality	potential	–
present conditional expressing irreality	irreal	–
present conditional expressing potentiality or irreality (unresolved)	potential|irreal	–
indicative expressing assertion	asserted	ant / sim / post
indicative instead of the conditional with the meaning of potentiality	asserted	ant / sim / post
indicative instead of the conditional with the meaning of irreality	irreal	ant / sim / post

Table 1. Rules for assigning the new grammateme factmod. In the 1st column, the mood of the verb form is given. According to the meaning of this form, the appropriate value of the factmod is chosen (the 2nd column). In the 3rd column, it is indicated whether one of the values of the grammateme tense is to be chosen (if not, – is filled in).

Concerning the present conditional, one of the values of the grammateme factmod is to be chosen by the annotator on the basis of context, knowledge of situation etc.: if this verb form expresses an irreal event, the value irreal is assigned to the corresponding node; if a potential event is concerned, the value potential is the right one; in case the annotator is not able to decide between the two interpretations, both values are to be filled in (potential|irreal). As for the expressions of irreality, the past conditional and the present conditional are thus considered as synonymous means: for sentences which differ just in the conditional form used, identical tectogrammatical representations are supposed.

Besides these basic cases, we also deal with tectogrammatical representation of the sentences in which an indicative verb form is used instead of the conditional one (cf. the results of the substitution test in Section 3). Examples such as the governing clause in the sentence (7), in which the conditional can be substituted for more than one indicative form (cf. (7')), indicate that these sentence pairs are semantically close to each other, though not synonymous. Therefore, they have to be represented differently at the tectogrammatical layer. Rules concerning the assignment of the indicative which occurs instead of the conditional are specified in the next subsection.

(7) *... jestliže bychom se rozhodli stát jadernou zemí,* ***musela by následovat*** *celá řada kroků ...* '... if we decide to become a nuclear country, a whole series of steps **would have to follow** ...' (PDT 2.0)
(7') *... jestliže bychom se rozhodli stát jadernou zemí,* ***musí následovat / bude muset následovat*** *celá řada kroků ...* '... if we decide to become a nuclear country, a whole series of steps **has to follow / will have to follow** ...'

5.3 Annotation of the indicative in originally conditional contexts

We propose to represent the indicative in sentences rendering irreal events in the same way as the past conditional (or the present conditional with irreal meaning), i.e., by a tectogrammatical node with the value irreal in the grammateme factmod; additionally, the node is assigned with the grammateme tense, in which the temporal characteristic of the indicative form is captured. Unlike the annotation of the indicative in sentences expressing irreal events, the value asserted is chosen in the grammateme factmod and an appropriate value of the grammateme tense is assigned if the indicative is used instead of the present conditional with a potential meaning.

The reason for a different annotation of the indicative in irreal and potential contexts is the disparity of conditions under which the indicative can be used instead of the conditional in these two context types. In sentences rendering irreal events, the conditional can be substituted for the indicative only under relatively strong conditions (if the irreality is clearly expressed by some other means) while the conditional and the indicative are mostly freely interchangeable in potential contexts. As mentioned above, this interchangeability is related to the semantic closeness of the potentiality of the conditional and the future meaning of the indicative. If we decided to represent the indicative similarly to the conditional in sentences expressing potential events (by the value potential in the grammateme factmod in combination with a value of the grammateme tense), it would mean at the same time that we have to resolve at each future indicative form whether it expresses a future or a potential event, which we consider as impossible in many cases. Assignment rules for the grammateme factmod are resumed in Table 1. Examples of verb forms assigned with the values of the grammateme factmod (and grammateme tense if necessary) are given in Table 2.

ex. nr.	verb form	gramm. factmod	gramm. tense	ex. nr.	verb form	gramm. factmod	gramm. tense
(1)	*by oslavil*	irreal	–	(5)	*skončil by*	irreal	–
(2)	*by bylo bývalo*	irreal	–	(5')	*skončí*	irreal	post
(3)	*by bylo*	potential	–	(6)	*by vypukla*	potential	–
(3')	*bude*	asserted	post	(7)	*musela by následovat*	potential	–
(4)	*by byla*	potential	–	(7')	*musí následovat*	asserted	sim
	by byla	potential	–		*bude muset následovat*	asserted	post
(4')	*bude*	asserted	post				
	bude	asserted	post				

Table 2. Verb forms from the example sentences (1) to (7') and corresponding values of the grammatemes factmod and tense which are to be assigned to the tectogrammatical nodes representing the verb forms in question. Only verb forms which were written in bold in the examples are included in the table. Verb forms from the sentences marked with a hash mark (i.e., (1'), (2'), and (6')) were not assigned.

6 Conclusions

The present paper focused on the Czech conditional mood when expressing hypothetical events (we talked about the primary function of this mood). It was illustrated with several examples that the conditional is a semantically relevant means in contemporary Czech although it could be substituted for the indicative in some of the studied contexts. The conditional in its primary function should thus be included in the tectogrammatical representation of PDT 2.0.

For this purpose, the grammateme factmod was proposed. Three values of this grammateme were suggested by means of which the difference between hypothetical and asserted events as well as the difference between two types of hypothetical events (the potential and irreal ones) are captured. Rules for manual assignment of the values were further described in the paper.

In the near future, rules for (at least) semi-automatic assignment of this grammateme have to be specified. Also tectogrammatical representation of other functions of the conditional and further modal means is to be elaborated.

Acknowledgements

I would like to thank Professor Jarmila Panevová for extensive discussion on the presented topic. I thank also Professor Eva Hajičová and Zdeněk Žabokrtský for valuable comments on the draft of this paper. The work reported on in this paper was supported by the project 1ET101120503.

References

[1] Bybee, J. L. (1985). *Morphology: A Study of the Relation between Meaning and Form*. Benjamins, Philadelphia.

[2] Daneš, F., Hlavsa, Z., Grepl, M. a kol. (1987). *Mluvnice češtiny 3*. Academia, Praha.

[3] Danilava, S. and Schommer, C. A Semi-Automatic Framework to Discover Epistemic Modalities in Scientific Articles.

[4] Gabrielatos, C. and McEnery, T. (2005). Epistemic Modality in MA Dissertations. *Lengua y Sociedad: Ivestigaciones recientes en lingüística aplicada. Lingüística y Filología*, 61:311–331.

[5] Grepl, M. a Karlík, P. (1998). *Skladba češtiny*. Votobia, Olomouc.

[6] Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and Ševčíková Razímová, M. (2006). *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia.

[7] Karlík, P. (1983). K významu kondicionálu minulého. *Slovo a slovesnost*, 44:12–21.

[8] Komárek, M., Kořenský, J., Petr, J., Veselková, J. a kol. (1986). *Mluvnice češtiny 2*. Academia, Praha.

[9] Osenova, P. and Simov, K. (2003). The Bulgarian HPSG Treebank: Specialization of the Annotation Scheme. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 129–140, Vaxjo, Sweden. Vaxjo University Press.

[10] Palmer, F. R. (2001). *Mood and Modality*. Cambridge University Press, Cambridge.

[11] Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31:71–106.

[12] Panevová, J., Benešová, E. a Sgall, P. (1971). *Čas a modalita v češtině*. Univerzita Karlova, Praha.

[13] Piqué, J., Posteguillo, S., and Andreu-Besó, J. V. (2001). A Pragmatic Analysis Framework for the Description of Modality Usage in Academic English Contexts. *Estudios de Lingüística Inglesa Aplicada*, 2:213–224.

[14] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

[15] Thompson, P., Venturi, G., McNaught, J., Montemagni, S., and Ananiadou, S. (2008). Categorising Modality in Biomedical Texts. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 31–34, Marrakech, Morocco. ELRA.

[16] Ševčíková, M. (2009). *Funkce kondicionálu z hlediska významové roviny. Disertační práce*. MFF UK, Praha.

[17] Šmilauer, V. (1966). *Novočeská skladba*. SPN, Praha.

The Creation of the Morphological Ambiguity Depository in Ukrainian

Olga Shypnivska[1] and Sergij Starykov[2]

National Taras Shevchenko University of Kyiv[1], Ukraine
Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukrainian[2], Ukraine

Abstract. Our paper deals with the morphological homonyms as a type of word sense disambiguation. The study provides a number of linguistic databases of morphological homonyms as a research environment. A brief description of a formal classification of the Ukrainian morphological homonyms according to different types and patterns, frequency dependence of these types from number of components and grammar properties are presented. The principles and structure of the electronic dictionary of morphological homonyms presented both in the dictionary of the word forms and in the stylistic differentiated texts corpus are described.

1 The morphological homonyms as a type of word sense disambiguation in Ukrainian. Some problems and proposed solutions

Ambiguity, the phenomenon that a word has more than one sense, poses difficulties for many current natural language processing (NLP) systems. The potential for ambiguous readings is completely unnoticed in normal text and flowing conversation. Words may be polysemous in principle, but in actual text with contextual knowledge and topical associations there is very little real ambiguity – to a person. In the computational linguistics word sense disambiguation is considered as the problem of computationally determining which “sense” of a word is activated by the use of the word in a particular context.

Ukrainian as many Slavic is very flexible. Ambiguity at the morphological level is very widespread. Analysis of the morphological tagging texts shows that more than 70 % word forms are morphological homonyms.

Consider the following sentence taken from the corpus, each containing 29 word forms:

За <Ь0Н0ПВПРПТ> результатами <ЙЮ> дослідження <ЛИЛРЛВЛАЛУ> видно <Х0>, що <Ь0СПМQMU> серед <КЕПР> так <СПССЬ0Н0КЕ> званих <АЕАУАЯ> простих <АЕАУАЯ> громадян <ЙЕЙУ> явно <Н0> проглядається <ГЯ> бажання <ЛИЛРЛВЛКЛАЛУЛШ> мати <ГФЙАЙУЙШКИКК> "одного <ЧРЧВЧЧОРОВОЧ> господаря <ЙРЙВ> в <ПВППППР> домі <ЙП>", який <ОИОВ> несе <ГЯ> всю <ОЛ> відповідальність <КИКВ> і <СПССN0Ь0> за <Ь0Н0ПВПРПТ> внутрішню <АЛ>, і <СПССN0Ь0> за <Ь0Н0ПВПРПТ> зовнішню <АЛ> політику <ЙДЙПЙККВ>.

The application of the morphological parser indicates that 20 word forms are morphologically ambiguous. For example, the word form *дослідження* means the nominative, the genitive, the accusative cases of the neuter noun in singular and the nominative, the accusative cases of the neuter noun in plural. The word form *мати* denotes two grammatical classes – the verb and the noun. *Мати* we can use as infinitive of the verb. *Мати* as a noun has different senses: the nominative, the vocative of the feminine noun in singular and the nominative, the accusative, the vocative of the masculine noun in plural. (The tagset consists of two-letter symbols that correspond word forms with a common set of morphological values. It is described in [9].

Current researches in Slavic, reported in the literature, provide some descriptions for the morphological ambiguity [3; 4]. There are a lot of works which represent attitudes and methods of its resolving [4; 5; 6; 7; 10]. But sometimes (in articles concerning inflexible languages) it argued that homographs are generally much easier to disambiguate than polysemous words [2]. Little researches have been devoted to Ukrainian [7]. These studies use rule-based approaches.

Each word relies on different kinds of knowledge for disambiguation. Senses of ambiguity words can be related in different ways. Therefore it is necessary to define all possible types of relations between morphologically ambiguous word forms. The results can be used to identify all possible patterns of disambiguation which expect valid rules their resolving.

The profusion of different types of homonyms induces us to create of the depositary of morphological ambiguity in Ukrainian language. In the paper we present a formal classification of the morphological homonyms according to different types and patterns. We describe a number of linguistic databases of morphological homonyms as a research environment and demonstrate the qualitative and quantitative description of types and patterns of the morphological homonyms in contemporary Ukrainian. Besides, the electronic dictionary of morphological homonyms is presented.

2 Principles, structure and the main goal of linguistic databases of morphological homonyms in Ukrainian

We consider the depositary of morphological ambiguity in Ukrainian as a research environment. These linguistic databases generally designed:

- to represent a list of potential morphological homonyms in Ukrainian both homonyms belonging to different part-of-speech and homonyms belonging to one part-of-speech
- to show types and patterns of the morphological homonyms and their occurrence probability in the real texts
- to possess all contexts for every pattern.

But the central question is how to form a complete list of potential morphological homonyms. Ukrainian is very flexible and there are a lot of open class words. Standard dictionary do not contain morphologically ambiguous word forms. There is a need for new, machine techniques and good collections.

As a base we used the dictionary of the word forms and the morphological parser. About the Ukrainian grammar dictionary and morphological parser see [9]. The dictionary has 3 094 174 word forms and 1 554 372 of them are morphologically ambiguous. From this list we selected 610 612 morphological homonyms.

The main aim of our attitude is to define possible differences among homonyms. We distinguish two kinds of homonyms: homonyms belonging to different part-of-speech (for our example – word *мати*) and homonyms belonging to one part-of-speech (in our examples – word *дослідження*). In the course of the research it was observed that only 2 % (10 215) of the list are homonyms belonging to different part-of-speech and 98 % (600 409) are homonyms belonging to one part-of-speech. The integration of homonyms with the same grammar senses allows us to determine 1 486 patterns from which 820 describe homonyms belonging to different part-of-speech.

Based on the list of patterns according to the part-of-speech meanings, we have selected 92 types of morphological homonyms (81 of them represent homonyms belonging to different part-of-speech). Table 1 displays a fragment of these linguistic databases.

Word form	Pattern	Type
мов	ГЬКЕСПгЬ	V-N-Prepos
ніж	ГЬКЕСПгЬ	V-N-Prepos
поверх	ЙВЙИН0ПР	N-Prepos-Adv
плач	ГЬЙВЙИ	V-N
моделі	КАКДКПКРКУКШ	N
фізика	ЙВЙРКИ	N

Table 1. The linguistic databases of morphological homonyms in Ukrainian

The first column provides morphological homonyms. Columns second and third provide patterns and types which correspond with them.

Linguistic information represented in this way can be combined in different manners according to your requirement. For example, we can take homonyms two of the most frequent patterns with the grammatical meaning of noun and verb. In this pattern verb is used in imperative and noun has a form of masculine. See Table 2 for examples.

Word form	Pattern	Type
голуб	ГЬЙВЙИ	V-N
вихри	ГЬЙАЙУЙШ	V-N
бубни	ГЬЙАЙУЙШ	V-N
плач	ГЬЙВЙИ	V-N
виїзди	ГЬЙАЙУЙШ	V-N
твори	ГЬЙАЙУЙШ	V-N

Table 2. The linguistic databases of morphological homonyms belonging to different part-of-speech in Ukrainian (Patterns)

The linguistic databases of morphological homonyms belonging to different part-of-speech can give use the most frequent types of homonyms in Ukrainian. Table 3 displays some basic statistics (the second column provides the quantity of homonyms).

Type	Amount	Word form
Adj-N	5682	вартові, виборні
V-N	2406	коти, лови
Adj-V	588	бадьорим, білим
V-Adv	389	значимо, мислимо
N-Adv	342	візаві, ажур
Adj-Adv	272	балакуче, захоплююче

Table 3. The linguistic databases of morphological homonyms belonging to different part-of-speech in Ukrainian (Types)

Verbs, nouns, adjectives are the most flexible and most open grammar classes in Ukrainian. For examples, in the Ukrainian grammar there are more then 76 000 nouns. That's why 6 two-component types (see Table 3) of homonyms with these part-of-speeches present 95 % of all types and types which have more then two components coverage only 5 % of homonyms belonging to different part-of-speech.

Studying the data, we observed that near 75 % of homonyms can have similar components in their meanings. Let us compare, for examples, homonym *плач* (see Table 2) it can be as verbal noun and verb with common components of meaning and homonym *мов* which can denote as different part-of-speeches (noun and preposition) and can differ number and gender for noun *мова*. We distinguished two kinds of types: complex semantic types and simple formal types.

2.1 Linguistic resources for homonyms functioning investigations in texts and tools for resolving their ambiguity

To appreciate the quantity of homonyms and their occurrence probability present in contemporary texts (fiction, journalistic genre, scientific texts), we processed the 3 000 000 word forms added to the database with potential types. This material shows

us that 62 % of potentional homonyms belonging to different part-of-speech were occurred at the texts. The most frequent types include unflexible part-of-speeches: conjunction, preposition, interjection. These types are not productive in the dictionary of the word forms and they are not very regular. Though there is clearly a need to explore the most regular types. As far as verb and noun are the most informative part-of-speeches the type Verb-Noun is very regular. For example, this type of homonyms makes 34 % of all homonyms in journalistic genre. To compare their regularity, types include unflexible part-of-speeches constitute only 3 %.

Table 4 displays some basic statistics for the most frequent types.

Type	Middle occurrence probability
Interj-Conj-Part	13,83
Conj-Part	7,89
Interj-Prepos-Part	7,39
N-Prepos	7,12
Interj-Prepos	5,75
N-Pron	5,64
Pron-Conj-Part	5,19

Type	Middle occurrence probability
V-N	5,05
Adj-N	3,84
Pron-Conj	3,64
N-Adv-Conj-Part	3,48
Adv-Prepos-Part	3,07
Pron-Part	2,35
N-Adv	2,28

Type	Middle occurrence probability
Adj-Adv	1,75
N-Pron-Adv	1,28
N-Pron	1,19
V-Prep	1,04
Adv-Conj-Part	1,02

Table 4. Middle occurrence probability of the homonyms belonging to different part-of-speech.

There are a lot of previous works which deal with the disambiguation resolving. They introduced us different methods such as knowledge-based, unsupervised corpus-based, supervised corpus-based. In general see Eneko Agirre [2]. As a rule they concern lexical disambiguation in inflexible languages [2]. As for Slavic all studies use rule-based approaches [5; 6; 7; 8]. But for Ukrainian there is a need finding attitudes and tools which allow getting results to a high degree of accuracy according to homonyms qualities and their frequency in our language.

To achieve a compromise between main directions we suggest linguistic databases of contexts. In our attitude we combine statistics and a method of rules. These linguistic resources suppose its application as combinations of technics for resolving morphological disambiguation.

The linguistic databases of contexts consist of two tables. The first one represents formal grammar contexts for each homonym. The second represents particular formal grammar contexts for each pattern. These linguistic databases were a two-step process. The morphological annotated texts served as the base of contexts. In step one, automatically we define formal grammar contexts for each homonym and compute them. The length of a context is 7 words in the right and in the left from a homonym. In step two, we manually determine the particular formal grammar context for each instance and compute them too. We extract relevant grammar, semantic features for a given homonym. Also we point the length and the location of these

determining features and the relevant meaning for each instance. In this case automatically we unite the same grammar contexts. Finally, the linguistic databases of contexts were integrated with the general linguistic databases of morphological homonyms. Table 5 displays a fragment of the linguistic databases of particular formal grammar contexts for each pattern.

Pattern	Particular formal grammar context (PFGC)	Relevant meaning	PFGC length in the left	PFGC length in the right	Comments
ГЬЙАЙУЙШ	<АААУ>	Noun	1	0	Adjective is coordinated with noun
ГЬЙАЙУЙШ	< ГЭ >	Noun	1	0	
ГЬЙАЙУЙШ	< Ь0 > <ММО> <КИ>	Verb	1	1	

Table 5. The linguistic databases of particular formal grammar context

The results show that more then 96,6 % of morphological disambiguation can be resolving in this way. It was examined on new texts. Statistical calculations verified that particular formal grammar context in the left can disambiguate more then 80 % of all instances. We must admit, when homonyms have similar components in their meanings ambiguity can be still present as shown by the examples below (see also table 2):

"*Хочеш* <ГСГХ> *наблизитися* <гФ> *до* <ЛАПР> *ідеалу* <ЙДЙР> -<> *плати* <ГЬЙАЙУЙШКАКРКУКШ> "; "*Реквієм* <ЙИЙВ> *над* <ПВПТ> *двома* <ЧЮ> *найвідомішими* <АЮ> *в* <ПВППРР> *світі* <КДКПЙП> *студіями* <КЮ> *колись* <Н0ГЬ> *і* <СПССN0Ь0> *плач* <ЙИЙВГЬ> *над* <ПВПТ> *втратою* <КТ> *національного* <АРАВАЧ> *надбання* <ЛИЛРЛВЛКЛАЛУЛШ> *за* <Ь0Н0ПВПРПТ> *повної* <АЗ> *байдужості* <КРКДКП> *держави* <КРКАКУКШ>").

In the second sentence the formal grammar context length is not enough for this homonym.

3 The electronic dictionary of morphological homonyms belonging to different part-of-speech

As an example of the created database application we offer the electronic dictionary of morphological homonyms belonging to different part-of-speech. For this time we know 2 dictionaries for Slavic which represent the morphological homonyms [1; 4]. Our dictionary has several advantages. It is machine readable and corpus-based. The electronic dictionary was automatically generated and can be integrate to different current natural language processing systems or to other dictionaries. It can be used to study many aspects of the morphological ambiquity.

The main goal of it is to represent morphological homonyms occurred both in the word forms dictionary and in the stylistic differentiated texts corpus.

It contains:

- a list of the potential and real morphological homonyms belonging to different part-of-speech;
- all types and patterns correspond with homonyms;
- word lemmas correspond with each homonyms;
- the occurrence probability of the morphological homonyms belonging to different part-of-speech present in contemporary texts (fiction, journalistic genre, scientific texts).

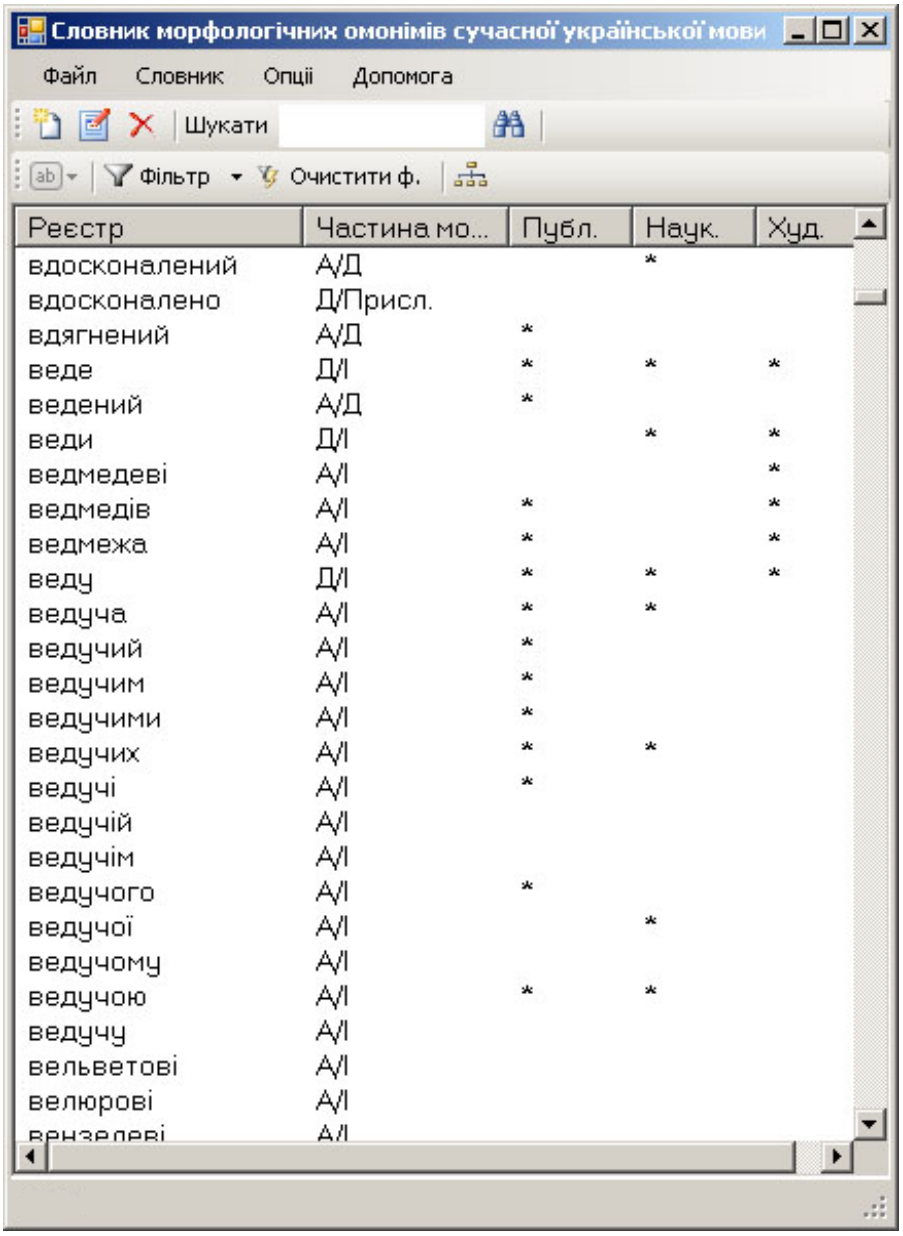
As far as the dictionary has machine readable format it can give us a lot of answers for different queries. For example we can take a list of word lemmas which can have homonyms. Table 6 and picture 1 show the format of this dictionary.

Homonyms	Pattern	Type	Word lemma
стелю	ГГКВ	V-N	стелити стеля
ніж	ГЬЙВЙИСП	V-N-Conj	ніж ніжити ніж

Table 6. The database of the electronic dictionary of morphological homonyms

4 Further developments

Our goal was to build universal, accessible language databases for Ukrainian in cooperation with everyone who wants to participate. The presented here language databases are now available for further researches and for making morphological disambiguation systems of Ukrainian homonyms. Besides, machine readable format of the morphological homonyms dictionary allows integrating it to different dictionaries in order to represent all semantic and functioning features of each word. Finally, obtained results can be used in comparative researches of Slavic.



Реєстр	Частина мо...	Публ.	Наук.	Худ.
вдосконалений	А/Д		*	
вдосконалено	Д/Присл.			
вдягнений	А/Д	*		
веде	Д/І	*	*	*
ведений	А/Д	*		
веди	Д/І		*	*
ведмедеві	А/І			*
ведмедів	А/І	*		*
ведмежа	А/І	*		*
веду	Д/І	*	*	*
ведуча	А/І	*	*	
ведучий	А/І	*		
ведучим	А/І	*		
ведучими	А/І	*		
ведучих	А/І	*	*	
ведучі	А/І	*		
ведучій	А/І			
ведучім	А/І			
ведучого	А/І	*		
ведучої	А/І		*	
ведучому	А/І			
ведучою	А/І	*	*	
ведучу	А/І			
вельветові	А/І			
велюрові	А/І			
вензелеві	А/І			

Fig. 1. The format of the electronic dictionary of morphological homonyms (fragment)

References

[1] D. Buttler editor. (1984), Słownik polskich form homonimicznych, Wrocław – Warszawa – Kraków – Gdańsk – Łódź, Wyd-wo Polskiej Akademii Nauk.

[2] E. Agirre and Philip Edmonds editor (2007), Word sense disambiguation. Algorithms and applications http://www.wsdbook.org/

[3] El. Awramiuk (1999), Systemowość polskiej homonimii międzyparadygmatycznej. Białystok: Wyd-wo Uni-tu w Białymstoku.

[4] J. G. Anoshkina (2001), Slovar omonimichnych slovoform russkogo jazyka, Moskva. http://www.irlras-cfrl.rema.ru/homoforms/

[5] J. V. Zinkina, N. V. Pjatkin, O. A. Nevzorova (2001), Razreshenije funkcionalnoj omonimiji v russkom jazyke na osnove kontekstnych pravil. // http://www.dialog-21

[6] Ł. Dębowski (2001), Tagovanie i dezambiguacija morfosyntaksyczna. Przegląd method i oprogramovania, Warszawa.

[7] T. A. Graznuchina, L. G. Bratyscczenko, N. P. Darchuk, V. I. Krytska, T. K. Puzdyreva, L. V. Orlova (1989), Shlachy unuknennja omonimiji w systemi automatychnogo morphologichnogo analizu [in:] Movoznawstvo, №5.

[8] T. J. Kobzareva, R. N. Afanaseva, (2001) Universalnyj modul predsintaksicheskogo analiza omonimiji chastej rechi w russkom jazyke na osnove slovara diagnosticheskich situacij. // http://www.dialog-22.ru

[9] V. A. Shyrokov editor (2005), Korpusna lingvistyka, Kyiv.

[10] M. Rudolf (2004), Metody automatychnej analizy korpusu tekstów polskich, Warszawa.

Frequency of Words and Their Forms in Contemporary Slovak Language Based on the Slovak National Corpus

Mária Šimková and Miroslav Ľos

Ľ. Štúr Institute of Linguistics,
Slovak Academy of Sciences, Bratislava, Slovakia

Abstract. Highest ranking words, word forms and n-grams based on their absolute frequencies are presented and compared with respect to the size, time of conception and stylistic focus of the source corpora. Corpus coverage by the most frequent words is discussed and distribution of words among morphological categories is analyzed. The dependencies of word use on style and genre of text are uncovered, utilizing the morphological tagging available in the Slovak National Corpus.

1 Introduction

The development of Quantitative (Statistical) Linguistics has seen more than 100 years since publication of the first European frequency dictionary (F. W. Kaeding: Häufigkeitswörterbuch der deutschen Sprache, 1897), and more than forty years since the first computer-aided compilation of language statistics from an electronic corpus was presented (H. Kučera & W. N. Francis: Computational Analysis of Present-Day American English, 1967). Many more languages have since been augmented with a frequency dictionary and/or statistical characteristic of their own, satisfying demand coming not only from disciplines such as NLP, (both monolingual and translation) lexicography, first and second language education, logopeadics, neurology, psychology etc., but also from language hobbyists interested in various quantitative properties of particular language units as well as the language itself. The existence of multiple voluminous corpora and the tools to process them allow for rapid collection of such statistics, often inspiring further research.

2 Word, word form and n-gram frequencies

J. Mistrík ([1] 1969, [2] 1985) manually analyzed word, word form and language construct frequencies in Slovak language, based on a text of 1 million proper words (i. e. without punctuation and any other non-word tokens). Some of his findings have been confirmed by data from the Slovak National Corpus[1], whose latest main corpus – prim-4.0 – made available in early 2009, counts about 550 million tokens. For example, the five lemmas (*a, v, na, sa, byť*) are consistently the top five, or among the top ten, highest-ranking lemmas across all its specializations, even in the spoken corpus.

[1] http://www.korpus.sk

Corpus	FSS 1969	prim1 2004	prim-4.0 2009	s-hovor 2009
Size	1 mil. w	200 mil. t	550 mil. t	434,676 t
1.	a	byť	byť	byť
2.	byť	v	a	to
3.	sa	a	v	a
4.	v	sa	sa	že
5.	na	na	na	sa
6.	on	ten	to	tak
7.	ten	ktorý	ktorý	ja
8.	že	s	s	v
9.	z	z	že	na
10.	ako	že	z	no

Table 1. Ten highest-ranking lemmas in general corpora and in the spoken corpus

Individual style of some authors, or smaller, specific texts, utilizing some characteristic language constructs (cf. Šimková 2008 [3]), could present statistics skewed from those obtained from general corpora. We created three corpora of distinct size and specialization: journalism (one month of the newspaper Smena), nonfiction (selection of texts regarding belief) and fiction (works of A. Habovštiak). The top rankings of lemmas from these corpora were essentially consistent with the general corpora.

Corpus	Smena 1968	belief	Habovštiak
Size	296,917 t	14,697,297 t	1,088,695 t
1.	a	a	sa
2.	v	byť	a
3.	byť	sa	byť
4.	sa	v	on
5.	na	na	na
6.	že	ktorý	v
7.	to	on	že
8.	ktorý	že	aj
9.	s	s	čo
10.	z	to	keď

Table 2. Ten highest-ranking lemmas in specialized corpora

Next to the top five lemmas, the corpora usually contain pronouns *ten* or *to*, *ktorý* (less often used as a conjunction), conjunction *že*, and prepositions *s*, *z*. Fiction works stand apart from other texts, for example in the higher use of the word *on* (he) both in the Habovštiak corpus and the corpus of Mistrík, mainly consisting of fiction.

Also notable is the higher use of the conjunction and particle *aj* (too, also, even) and the diminished conjunctive role of the pronoun *ktorý* (which), that has likely been partially replaced by the pronouns *čo* (that) and even *keď* (as, when), suggesting here a fundamental difference in fiction vs. nonfiction authors' selection of sentence structures.

The four most frequent words in the corpus prim-4.0 after the verb *byť* (to be) are also present in the top ten of both bigrams (*a* one time, 3× *v*, 5× *sa* and 2× *na*) and trigrams (3× *v*, 2× *na* and *sa*). These top-tens also contain the most frequent substantive *rok* (year), ranking 9th among bigrams and 3rd or 4th (ranked by *na druhej strane* if counting ignoring character case) among trigrams. Analysis of the various texts shows that time units in general (e. g. *rok*, *týždeň*, *deň*, *hodina*) have substantial frequencies in all works independent of the work's style and genre.

Notable among the most frequent trigrams are secondary prepositions (ranking 2nd and 4th, respectively), which is likely a result of the high proportion of journalistic texts in the subcorpus prim-4.0-public, the source of the list of bi- and trigrams in table 3.

Rank	Bigram	Frequency	Trigram	Frequency
1.	nie je	392,383	že je to	35,800
2.	sa v	315,831	v súvislosti s	35,021
3.	sa na	298,802	v tomto roku	32,553
4.	je to	295,003	v porovnaní s	27,489
5.	že sa	282,920	to nie je	26,740
6.	som sa	277,091	Na druhej strane	26,255
7.	by sa	258,065	na druhej strane	23,464
8.	a v	245,519	by som sa	21,807
9.	v roku	235,353	Nie je to	20,509
10.	na to	203,130	že by sa	20,411

Table 3. Ten highest-ranking bigrams and trigrams in prim-4.0-public

Whenever corpora differ in size by multiple orders of magnitude, the statistics collected on them naturally diverge in several parameters, such as the accumulated share of top-ranking lemmas (see Table 4).

Ranks	FSS	prim-4.0	Difference
1. – 10.	18.6 %	17.98 %	- 0.62
11. – 20.	6.4 %	5.74 %	- 0.66
1. – 20.	25.0 %	23.72 %	- 1.28
21. – 30.	4.0 %	3.51 %	- 0.49
1. – 30.	29.0 %	27.23 %	- 1.77
1. – 100.	41.5 %	37.81 %	- 3.69

Table 4. Corpus coverage by most-frequent words (excluding punctuation)

While according to J. Mistrík's Frequency Dictionary of Slovak Language (1969), the top one hundred words cover 41.5 % of the text (when discounting random or rare words with zero dispersion, this value rises to 56.13 %), this coverage is almost 4 % lower in prim-4.0. This can be explained by the rising number of random tokens as a corpus grows in size. It is further supported by lemmas from prim-4.0 with frequencies 1 or 2, which represent almost 66.5 % of all lemmas in the corpus, of which an estimated half are numerals, typos, words in foreign scripts or other symbols.

3 Frequencies of morphological categories

In the corpora based on written text, the first two placed are held by substantives and verbs (see Tab. 5). The 3rd place is held by prepositions in contemporary texts, in contrast with pronouns in the corpus of Mistrík, which may signify a higher level of abstraction in the discourse, perhaps as a consequence of the higher share of journalistic and professional works in prim-4.0. The ranks of adjectives, conjunctions, numerals and interjections is the same between these corpora; the ranks of the latter two are also shared with the spoken corpus. The generally low share of interjections in all corpora is nonetheless much higher in the spoken corpus (0.31 %) than in prim-4.0 (0.058 %). As was to be expected, verbs and pronouns take the first two places of the spoken corpus; moreover, the share of conjunctions and particles is quite higher. The statistics on adverbs shows across all corpora, that their potentiality (being derivable from both adjectives and substantives) in the language system is much higher than actual use.

Corpus	FSS 1969	prim-4.0 2009	s-hovor 2009
Size	1 mil. w	550 mil. t	434,676 t
1.	substantives	substantives	verbs
2.	verbs	verbs	pronouns
3.	pronouns	prepositions	substantives
4.	adjectives	adjectives	conjunctions
5.	prepositions	pronouns	particles
6.	conjunctions	conjunctions	prepositions
7.	adverbs	particles	adjectives
8.	particles	adverbs	adverbs
9.	numerals	numerals	numerals
10.	interjections	interjections	interjections

Table 5. Frequency of words by morphological category

The highest share of substantives classified by gender belongs to the masculine. Since the morphological annotation of the SNC distinguishes between animate and inanimate masculines (see `http://korpus.juls.savba.sk/usage/morpho/`), the order of genders thus split is instead as follows: 1st the feminine, 2nd inanimate masculine (i), 3rd animate masculine, and 4th the neuter. This order is consistent between works' styles and genres, with the exception of professional text (prf), having the animate masculine gender last, likely due to the generally abstract and impersonal style of such text (see Fig. 1).

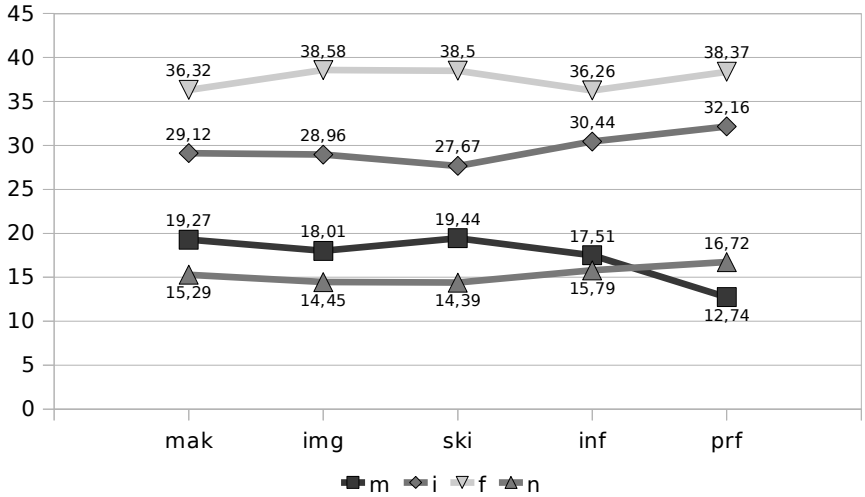


Fig. 1. Distribution of substantives by gender (%)

The manually morphologically annotated corpus (mak), used as reference in present analysis due to its negligible error rate, is closest in its gender distribution to the subcorpora of journalistic and fiction/artistic text (tagged inf and img, respectively). This reflects the bigger share these two types of texts have in mak (of the 1 207 939 tokens, 44.3 % are journalistic, 36.7 % fiction, and 19.0 % professional). The subcorpus ski contains all original text from img created by Slovak authors. Comparing it to its parent corpus, we see the same share for both feminine and neuter substantives, but a shift of some 1.3 percentage points in favor of animate masculines in the Slovak fiction corpus.

The distribution of substantives by number and case (Fig. 2) shows a clear preference towards singular forms, holding top four ranks in all main four corpora analyzed. The natural first is Nsg, the common 4th is Lsg. The second and third places belong to Gsg and Asg, with not so small differences ranging from 2.5 % points in mak to more than 7 % points in prf. In fiction, the accusative wins due to the higher usage of verb-object phrases. The journalistic and professional texts contain more substantives in genitive, which matches the higher use of genitive attributes in noun phrases.

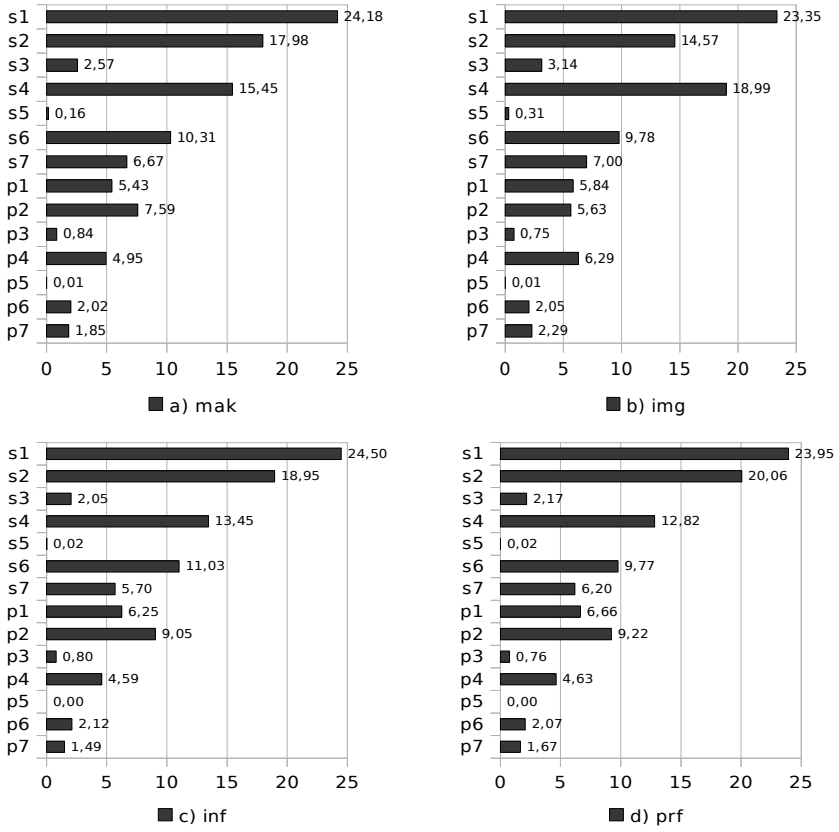


Fig. 2. Distribution of substantives by number and case (%)

Of plural forms, the genitive enjoys the highest usage (i. e. 5th place overall), most visibly again in journalistic and professional texts, followed in these by Npl by a margin of 3 % points. In fiction, Isg holds the 5th place, then followed by Npl and Apl. Gpl ranks 8th in fiction. The rarest case is dative, both in singular (9th overall, 10th in pub) and in plural (12th).

Places 9th–11th, containing Dsg, Lpl and Ipl., have only small differences among each other (0.5 % points in pub and prf to about 1 % point in img). A category of its own belongs to the vocative, regarded as just a relic in contemporary Slovak language system due to its near absolute homonymy with nominative. The few exceptions (e. g. *otče, majstre, bratu, priateľu*) represent a negligible part of the corpus, mostly only present in fiction (0.31 % in sg and 0.01 % in pl).

The distribution of prepositions by case (Fig. 3) is naturally distinct from that of the substantives, as the most frequent nominative completely lacks prepositions, and both genitive and accusative are often used without any preposition. The first place is thus taken by locative prepositions, mainly used by journalistic and professional works. The frequencies of genitive and accusative prepositions, chiefly used in fiction, are rather close to each other. Fiction works also use instrumental and dative prepositions more heavily, to express predicate complements.

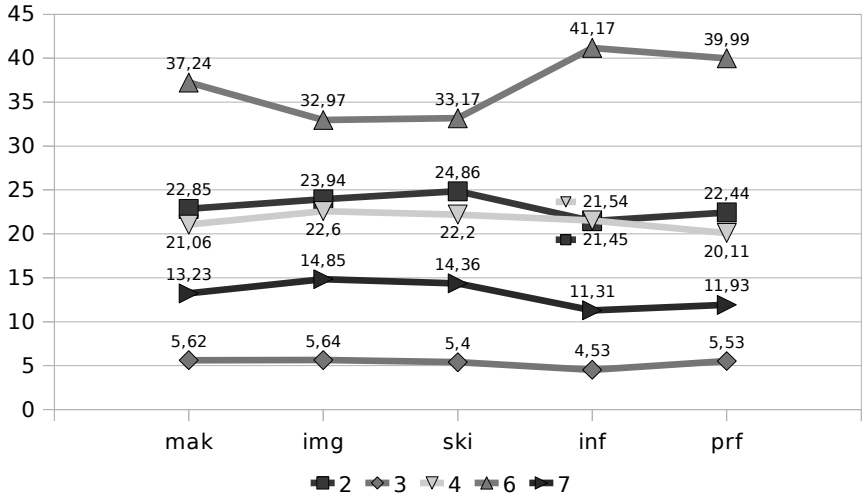


Fig. 3. Distribution of prepositions by case (%)

From the basic verb forms recognized by the morphological annotation in the SNC (i. e. B – future of *byť*, H – gerund, I – infinitive, K – indicative, L – l-participle, M – imperative), the l-participle (i. e. past tense and conditionals), the present indicative and the infinitive are the most frequent (Fig. 4). The other three forms are very specific, with only little applicability. The indicative is the most frequent in journalistic and professional text (represented by up to one half of all verbs in the subcorpus), while the l-participle is used the most in fiction (again nearly 50 % of all verbs).

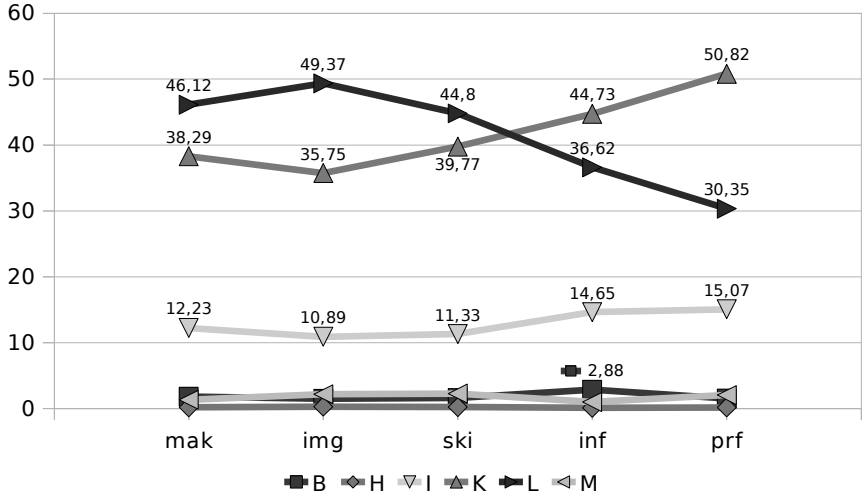


Fig. 4. Distribution of verbs by form (%)

The division of verbs by person-agreement is dominated by the third person singular (Fig. 5), having more than 50 % share across all corpora except Slovak fiction, where they are just below. In fiction, the second place is held by 1st person sg, which drops sharply in other styles towards almost zero in professional text. Journalistic and professional text has instead established 3rd person pl. as the second. Considerable is also the share of the indeterminate verb forms (wrt. number and person), marked under '--' in the chart, where the person is concealed, e. g. due to omitted subject pronoun (e. g. *Prišiel*. – ja, ty, on?; *Prišli*. – my, vy, oni?).

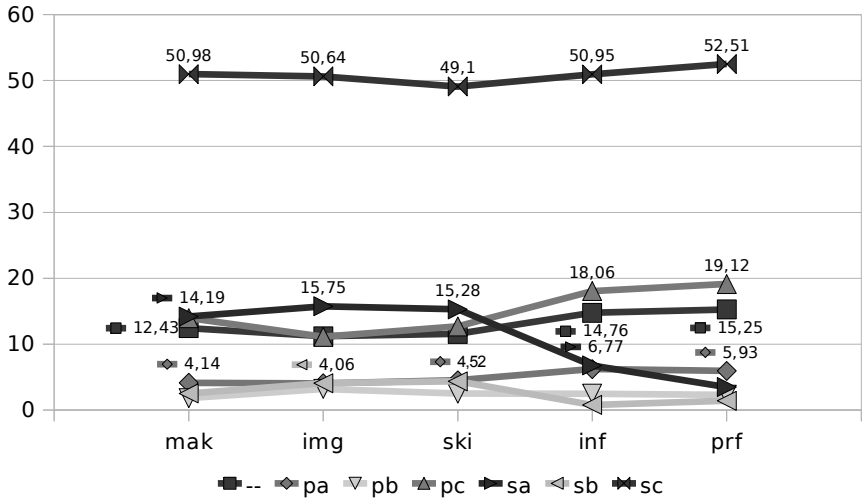


Fig. 5. Distribution of verbs by number and person (%)

4 Conclusion

After six years of building the primary database of Slovak National Corpus (made of texts originating in years 1955 – 2009), its quantitative (550 million tokens), as well as qualitative (i. e. style and genre structure, lemmatization and morphological tagging) properties allow for quantitative analysis with significant results. Much data can also be compared to observations of Slovak language made available 40 years ago, or analogous latest results from other languages.

Linguistic interpretation of frequency-based indicators of select lexical categories and word forms present in contemporary Slovak language presents one of the first contributions to this discipline based on the SNC. We have researched absolute frequencies of presented phenomena based on all written text in the corpus, as well as their distribution among the three main styles: fiction/artistic, journalistic and professional. Even the partial results published in present paper signal a marked tendency of the journalistic and professional styles to merge. We also contrasted some of the results with data from the early Slovak Spoken Corpus, which is too small at the moment to have reliable conclusions drawn from.

Detailed information on frequency-based parameters of contemporary Slovak language will be published in the Frequency dictionary currently in preparation at the Slovak National Corpus.

References

[1] Mistrík, Jozef: Frekvencia slov v slovenčine. Bratislava: Vydavateľstvo SAV 1969. 728 p.

[2] Mistrík, Jozef: Frekvencia tvarov a konštrukcií v slovenčine. Bratislava: Veda 1985. 320 p.

[3] Šimková, Mária: Jazykové prostriedky vo vybraných dielach Martina Rázusa (Analýza na báze textov Slovenského národného korpusu). In: Martin Rázus – politik, spisovateľ a cirkevný činiteľ. Ed. M. Pekník. Bratislava: Veda 2008, p. 238 – 249.

Analysis of the Means Expressing Strong 'Necessity Not To' in English and Czech Based on General and Parallel Corpora

Renata Šimůnková

Technical University of Liberec, Czech Republic
renata.simunkova@tul.cz

Abstract. The paper presents results of a corpus-based study of basic means used to express strong deontic 'necessity not to' in English and their counterparts in Czech. In particular, by means of co-text analysis different semantic values of these means of expression are delimited. It is further studied to what extent the choice of particular means is influenced by co-occurring means at different levels: lexical (collocations), grammatical (colligations), pragmatic, as well as by the type of discourse.
The matter is first studied within one language and then the structures of semantic field of 'necessity not to' in English and in Czech are compared.

1 Introduction

The initial impulse for this corpus-based study resulted from the work on a dissertation thesis whose aim was to compare means used to express necessity in English and Czech based on contemporary fiction and the corresponding published translations. It has been discovered that the difference in meaning between CANNOT and MUST NOT is by no means clear and that the generally assumed correlation between English MUST NOT and Czech NESMĚT, and English CANNOT and Czech NEMOCI is not always valid. The scope of the study, however, did not provide enough material to allow the deduction of any general conclusions.

The principal goal of this project is, therefore, to study the basic means used to express strong 'necessity not to' with the use of corpora in both English and Czech in the following steps:

- collection of the relevant means from the Czech National Corpus and British National Corpus
- classification of the means according to their different interpretations based on concrete criteria (e.g. objective/subjective modality, a ban/moral impropriety, etc)
- study of the reasons for the choice of particular means based on the study of co-text (collocations, grammatical, pragmatic and stylistic aspects)
- contrastive study of the choice of corresponding means in the two languages based on the parallel corpus Intercorp

2 Modality

The concept of modality is universal; it is not found only in English or Czech, but in the majority of, if not in all, languages. According to Palmer (1986: 7) "it is probable that there are very few languages that do not have some kind of grammatical system of modality." It represents an extensive and complex problem and since the fundamental means of its expression differ from language to language, the starting points of its description can vary between languages.

When trying to define modality in general, the keyword is *relation*: the relation of the speaker to the utterance, to the factuality and actualisation, the relation of the utterance to reality, the relation of the real world to possible worlds. Modality can be best seen as opposed to factuality – it states the content of the utterance not as a fact but as a potential fact dependent on certain conditions e.g. the authority and approach of the speaker (*You must take the exam now; you can take the exam no)*, the reliability of their judgement (*He must have forgotten about the meeting*).

Two basic concepts or notions of modality are generally distinguished – those of possibility and necessity. Each of these two concepts, necessity and possibility, is then further classified into kinds of modality – deontic, epistemic and often also dynamic.[1]

One additional feature of modality, which also appears in many works dealing with modality, is the strength of modality. The strengths are referred to by different names by different linguists e.g. Huddleston and Pullum (2002: 175) talk about "weak"," medium" and "strong", while Halliday (1991: 182) refers to the same concepts as "high", "median" and "low" yet they describe the same matter. On the basis of strength, the fundamental concepts of *obligation* and *high probability* on the "high" end of the scale and of *possibility* and *permission* on the "low" one are distinguished. Medium modality is somewhere in between, "though intuitively closer to the strong end than to the weak" (Huddleston and Pullum 2002: 177).

The basic distinction between deontic and epistemic modality is that between actualisation and factuality. Deontic modality has its illocutionary force and an utterance including deontic modality has the potential (another important key term connected with modality) to result in a particular physical act of doing or not doing something. An utterance including epistemic modality, on the other hand, has the potential of being/not being true. Halliday (1991: 183) describes deontic modality as calibrating the area of meaning between *Do it!* and *Don't do it!,* whereas epistemic modality as calibrating the area of meaning which lies between *Yes* and *No*.

An additional kind of modality can be distinguished – dynamic modality. For dynamic modality the key words are properties (of a situation) or disposition (of a person). Dynamic modality can be, therefore, seen as more objective than the two

[1] Some linguists (F. Kiefer, 1994) distinguish a greater number of different kinds of modality, but, for our purposes, such a detailed classification is not necessary. (With the pre-set aim in mind it would even be counterproductive.)

previous kinds. Since, however, a common practice is to distinguish only two basic kinds of modality such an approach is also adopted in this paper.

The distinction of the area of modality into the two afore-mentioned kinds is not absolute. In many cases it is a question of gradience rather than of clearly cut boundaries. The kind of interpretation is usually dependent on context because many modal expressions can be used with both deontic and epistemic interpretations.

3 Inescapable 'necessity not to' in English and Czech – theoretical foundations

Grammar books do not seem to deal in any big detail with the way in which 'necessity not to' is expressed. It is difficult to find theoretical studies of the differences in the meaning and use between the individual means. The main difference is the difference between MUST on the one hand and a set of negative modal means such as CANNOT, MAY NOT, BE NOT ALLOWED on the other hand, which is described as the difference between 'necessity not to do' and negative 'permission to do' = a ban. The former is felt stronger. This assumption is commented on by Leech (2004), and Palmer (1979): "Both these statements (with *may not* and *mustn't*) are prohibitions, but differ in that the second sounds rather more forceful, positively forbidding instead of negatively withholding permission" (Leech 2004: 95), "But there is an obvious difference between refusing permission (may not/can't) and laying an obligation not to (mustn't). With the former it is to be assumed that permission is normally required, while with the latter the speaker takes a positive step in preventing the action for which permission may not normally be required"(Palmer 1979: 65).

In Czech there are two basic modal verbs used: NESMĚT and NEMOCI, which since they are both used with external negation express negative permission. The only formal possibility in Czech to express the meaning of MUSTN'T is MUSET with the negative infinitive which is rare and restricted. Based on a short comment in Dušková (1994: 192): "Obvyklý zápor k dispozičnímu *may* je *must not*, popř. *can't*. Např. *you mustn't worry* nesmíte si dělat starosti, *we can't take the dog into the hotel* nemůžeme (nesmíme) vzít psa do hotelu. *One mustn't be proud*. Člověk nesmí být pyšný."[2], one might arrive at the conclusion that MUSTN'T corresponds to Czech NESMĚT and CAN'T to Czech NEMOCI. This might be further supported by looking into dictionaries which either do not deal with the issue at all or suggest the same correlation, e.g. in Lingea Lexicon NESMĚT = MUST NOT, MAY NOT. The situation, however, may not be so easy. Grammatically, CANNOT should be close in meaning to MAY NOT since they both take external negation. One clear distinction of MAY NOT from the other modal verbs used to express 'necessity not to' is the

[2] The usual negation to deontic may is *must not*, or *can't,* e.g. *you mustn't worry* nesmíte si dělat starosti, *we can't take the dog into the hotel* nemůžeme (nesmíme) vzít psa do hotelu. *One mustn't be proud*. Člověk nesmí být pyšný.

formal nature of MAY NOT. "Zápor *may not* (v dispozičním významu „nesmět") se vyskytuje v úředním jazyce" (Dušková 1994: 191)[3].

As reference materials did not provide any further information, native speakers of British English were asked for help. The opinions they provided were interesting, and surprising at the same time. They agreed on the assumption that while there exists a rather big overlap between MUST NOT and CANNOT, partly depending on a particular context and intonation, there is a basic difference felt between the two means, which does not seem primarily to reside in the force of the modal means as Leech (2004) suggested, but rather in the meaning. While CANNOT expresses either inability (which is outside the scope of this paper) or something which is officially (often by law) not allowed/banned, MUST NOT is used to express more personal involvement of the speaker describing something which is considered morally wrong. It can be seen as more or less in agreement with Palmer (1971) – (see the paragraph above). If something is usually not allowed, then one, if they in spite of this want or need to do it, has to ask for permission. It is, however, not logical to ask for permission to do something which is considered morally inappropriate. Such things are usually done either out of ignorance, or because one cannot help it.

Reference material on 'necessity not to' in Czech does not offer much information either. It was, therefore, attempted to deduce the negative meanings of the two modal verbs (NESMĚT, NEMOCI) from the positive ones. From Benešova's (1971: 132) classification of modal verbs it follows that MOCI is used to express possibility in general – it means in cases when the source of modality are outer circumstances as well as if the source of modality is a concrete human being and both in the case when the source of modality is identical with the source of action and when the source of modality is different from the source of action. SMĚT is used only if the source of modality is a concrete person and only in cases when the source of modality is different from the source of action. Based on these findings NESMĚT seems mainly to expresses things which are not allowed, which are officially banned. NEMOCI, on the other hand, refers, in addition to inability and impossibility based on outer circumstances, to things seen as bad, something that a kind of inner control prevents a person from doing e.g. – *to jí nemohu udělat, nemohu ji zradit*. According to this explanation, therefore, the correspondence between MUST NOT, CAN NOT on the one hand, and NESMĚT, NEMOCI on the other should be completely opposite. An example from Rowling's Harry Potter might be used to support this view:

> *They* ***cannot keep*** *the objects longer than that unless they can prove they are dangerous. (= it is banned by the Ministry) –* ***Nesmějí si*** *věci z pozůstalosti* ***ponechávat*** *déle, neprokáží-li …*

Since, however, in the same way as in English, also in Czech there is a big overlap between the two means, there is certain flexibility in their use.

---

[3] Negative *may not* (in deontic sence: nesmět) is used in official language

The comparison becomes simpler when the discussed modal verbs are used with past reference because due to formal reasons MUST NOT cannot be used. MUST does not have the past form and in order to express strong obligation with past reference other forms must be used, the most common being HAVE TO. HAVE NOT TO or more precisely HAD NOT TO, however, does not express an *obligation not to*, but a *possibility not to*. Apart from the just mentioned formal reason, there might be a semantic or logical reason for the non-existence of the past form of MUST with deontic interpretation. Since MUST, and in the same way probably also MUST NOT, is strongly subjective and being used when the source of modality is the speaker, it is highly illogical with past reference (even more so in its negative form). MUST NOT expresses something seen as wrong by the speaker, so if the speaker is at the same time the source of the action, the logical result has to be the non-realization of the action (no space for MUSTN'T here) or the action is accidental and then other means is preferable, e.g. *I did not want, mean, intend*… . It is also possible to speak about a ban in the past and then COULD NOT, BE NOT ALLOWED, BE NOT TO etc. is used.

4 Data collection

'Necessity not to' was studied on three corpora: Czech National Corpus (ČNK), British National Corpus (BNC) and parallel corpus Intercorp. Since the British National Corpus includes texts mainly from the 1990s, in Czech National Corpus part SYN2000 was chosen as the corresponding source. The corpus Intercorp, since it is comparatively smaller, was not reduced what concerns dates of publishing of the texts, since the issue studied is not likely to be significantly influenced by the time difference in the order of decades. However only works of fiction were selected and so was done also in the case of the other two corpora. The individual means studied were selected on the basis of the above mentioned dissertation. For English the following means were included: MUST NOT, CANNOT, COULD NOT, MAY NOT, NOT ALLOWED TO, NOT SUPPOSED TO, BE NOT TO. For Czech NESMĚT and NEMOCI were studied. In the case of NEMOCI and negative forms of CAN and MAY which apart from 'necessity not to' express also other meanings (which are in all the cases considerably more frequent) only cases where these means clearly expressed 'strong obligation not to' were included.

In the cases when the corpora provided a large number of tokens, 250 examples were processed. In the cases when less than 250 examples were available or manageable always at least 50 relevant cases were included in the study, with the exception of MAY NOT for which from 411 cases only 19 were relevant.

The whole issue was first studied for each language separately. The means found in the corpora were classified into categories on the following criteria: the source of modality and the source of action (on this basis modality is classified into subjective or objective) and within each category further two categories were distinguished: moral impropriety and a ban.

5 Results of the corpus-based study

The results gathered in the way described above are first presented in tables and then commented on. Table 1 presents data gathered from BNC, Table 2 from CNK.

Modal means	SUBJECTIVE		OBJECTIVE	
	Moral impropr. (in %)	Ban (in %)	Moral impropr. (in %)	Ban (in %)
MUST NOT	71	12	7	10
CANNOT	57		17	26
COULD NOT	78			22
MAY NOT		60		30
NOT ALLOWED				100
BE NOT TO	11	11	12	66
NOT SUPPOSED	8		44	48

Table 1. Classification of Means from BNC

Modal means	SUBJECTIVE		OBJECTIVE	
	Moral impropr. (in %)	Ban (in %)	Moral impropr. (in %)	Ban (in %)
NESMĚT	34	19		47
NEMOCI	28		41	31

Table 2. Classification of Means from ČNK

The data from BNC confirm that MUST NOT expresses mainly subjective modality, which is generally accepted truth and they also confirm the assumption of the native speakers discussed above that MUST NOT is predominantly connected with activities seen as morally inappropriate. Rather surprising are the findings connected with the modal verb CAN. In both present and past forms it mainly expresses subjective modality and moral impropriety. The higher percentage of moral impropriety meanings connected with the past form of CAN can be explained by the fact that since MUST NOT, which is supposed to be a chief means to express moral impropriety, does not have a past form and COULD NOT functions here as a suppletive form.

Another interesting fact which can be inferred from the data is the finding that for a ban English often uses other means than modal verbs. NOT ALLOWED TO, NOT SUPPOSED TO and NOT BE TO all predominantly express objective modality and a ban.

The relative closeness of meanings of MUST NOT and CANNOT/COULD NOT when expressing ‘necessity not to’ is further confirmed when the right-hand collocates of the two forms are studied. Among the 30 most frequent right-hand collocates of MUST NOT 16

were also among the most frequent collocates for CAN/COULD NOT. These were verbs *let, refuse, blame, complain, tell, say, get, risk, leave, allow, accept, pretend, disagree, ignore*. All these activities can rather be seen as morally inappropriate since they are not generally permitted or banned.

The issue was then studied on the parallel corpus Intercorp. English forms served as the starting forms in the search and their corresponding Czech translations were excerpted as well. The results are summarized in Table 3.

English modal means	SUBJECTIVE		OBJECTIVE	
	Moral impropr. (in %)	Ban (in %)	Moral impropr. (in %)	Ban (in %)
Must not	Nesmět – 74 no C (means no explicit modal means in Czech) – 5		nesmět – 21	
Can not	nemoci – 38 nesmět – 6 nelze upřít – 6		nesmět – 13	nesmět – 37
Could not	nemoci – 60 nesmět – 12 no C – 4	nesmět – 2 nemoci – 1	nemoci – 3 nesmět – 1 zakázat – 1	nemoci – 9 nesmět – 7
Not allowed				nesmět – 80 nedovolit – 20
May not				nesmět – 100
Not supposed	nemít – 16	nemít – 21	nemít – 10 nesmět – 11 no C – 5	nemít – 21 nesmět – 16
Be not to	nemoci – 7	nedovolit – 7 imperrative - 7	nemít – 7	mít zakázáno – 43 nesmět – 29

Table 3. Means from Intercorp

In the distribution of basic interpretations of English means expressing 'necessity not to' the data from the Intercorp correspond to those gained from BNC. As far as their translations into Czech are concerned, MUST NOT is mainly translated as NESMĚT. NESMĚT is, however, also the most frequent translation for CAN NOT. The data from Intercorp show that objective 'necessity not to' is in English frequently expressed by other means than modal verbs. This claim is further supported by an additional search in which English means used to translate Czech NESMĚT were excerpted with the following results: imperative construction – 15%, infinitive construction – 15%, would not – 9%, not want to – 8%, no C – 8%, not permit – 5%. All corpora also suggest close correspondence between MUST NOT and NEMOCI

since both these modal means are frequently used to express moral impropriety. This finding is further supported by the study of right-hand collocates of NEMOCI where the means expressing activities which are unlikely to be banned such as *tvrdit, nechat, zklamat, říci, upírat, přesvědčovat, souhlasit, žádat, kárat etc*[4] prevail.

Apart from MAY NOT, which is found mainly in formal academic or legal style and therefore was highly infrequent in the corpora used (only 19 relevant examples in BNC and in Intercorp it was found only in one text which was an example of academic style), the other studied means do not seem stylistically marked.

6 Conclusion

Corpora enable to study certain features of language on a large amount of material which was not possible or extremely time consuming in the past and thus to distinguish delicate shades of meaning. The study presented here is an attempt at such a delimitation of different meanings of means used to express 'necessity not to'. The generally assumed distinction between MUST NOT and CAN NOT residing in the fact that MUST NOT mainly expresses strong prohibition was not confirmed. Strong prohibition as such seems rarely to be expressed at all, the main interpretation of MUST NOT being subjective moral impropriety. The assumption that CANNOT unlike MUST NOT often expresses activities which require permission and therefore can be banned is valid only to a certain extent since the predominant meaning of CANNOT is also to express moral impropriety, although, in contradiction to MUST NOT, it expresses objective modality more frequently.

Learners and non-native users of both English and Czech should be aware of the oversimplification of the correspondence between MUST NOT and NESMĚT on the one hand and CANNOT and NEMOCI on the other and choose the corresponding means carefully and from a larger set of the means available.

Acknowledgement

This research has been supported by the Inner Grant Competition of FP TUL, grant IGS 69/2009.

References

Benešová, E., J. Panevová, and P. Sgall. (1971) *Čas a modalita v češtině*, Praha: Universita Karlova

Čermák, F., R. Blatná. (2005) *Jak využívat Český národní korpus. Praha*: Vivas.

Dušková, L. at al (1994) *Mluvnice současné angličtiny na pozadí češtiny*, Praha: Academia

Halliday, M., A., K. (1991) *An Introduction to Functional Grammar*, London: Arnold

[4] Claim, let, dissapoint, say, deny, persuade, agree, ask, reproach

Huddleson, R. and G. K. Pullum (2002) *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press

Hunston, S. (2002) *Corpora in Applied linguistics*. Cambridge: Cambridge University Press

Leech, G. (2004) *Meaning and the English Verb*, London: Longman

Palmer, F., R. (1979) *Modality and the English Verb*, London: Longman

Palmer, F., R. (1986) *Mood and Modality*, Cambridge: Cambridge University Press

Palmer, F., R. (1995) Negation and the modals of possibility and necessity, in *Modality in Grammar and Discourse*, edd Bybee J. and S. Fleischman, Amsterdam/Philadelphia: John Benjamins, 454 – 471

Research sources

Czech National Corpus: SYN 2000. Prague, 2000. Available on-line from `http://ucnk.ff.cuni.cz`

Czech National Corpus: Intercorp. Prague, 2006. Available on-line from `http://ucnk.ff.cuni.cz`

British National Corpus. XML edition. Oxford, 2007.

Diatheses in the Czech Valency Lexicon PDT-Vallex

Zdeňka Urešová and Petr Pajas

Charles University in Prague, UFAL MFF UK, Czech Republic

Abstract. An important design element in all lexicons, whether human-oriented or designed for computer processing, is the variability of forms in which lexical units described in the lexicon entries can occur in natural language utterances. If all such forms and variations were to be listed independently in the lexicon, its size would be enormous and it would be hard to maintain (every change would have to be copied to many entries). These problems can even multiply in the case of lexicons for computerized natural language applications, where entries must be explicitly and formally described in full detail.
As an inherent part of the Prague Dependency Treebank project ([9]; for its theoretical background, see the work of Sgall et al. [33]) a valency lexicon called PDT-Vallex ([10], [39], [40]) has been created and is publicly available, with over 8800 verb senses and their corresponding valency frames, linked fully to the treebank.
When a particular verb sense is used in a diathetic expression (passive construction, reciprocity, resultative or dispositional modality etc.), the surface expression of verb complements also changes ([40]). While the basic form "transformations" are well known, it is less obvious how to describe them for all the modalities, especially for the purposes of computer processing, where everything must be explicitly stated. We have found that these transformations can be described by a set of rules, which then allow to keep only a canonical (i.e., the active-voice) valency frame in the lexicon entry, and use these rules to obtain surface expression constraints for all the diatheses covered. This formalization have been used in the formal checking of the Prague Dependency Treebank project and it is used in other current projects as well.

1 Valency

Before we will concentrate on diatheses in the PDT-Vallex, let us make a little digression into the very notion of valency and diathesis.

1.1 Valency in general

This introductory section reviews some very basic facts about valency. Most writers on the subject cite Tesnière [38] as the one responsible for introducing the term of valency into modern linguistics. Tesnière uses the term valency for syntactic analysis of a sentence, so it was linked also to dependency. Active valency and passive valency are occasionally distinguished in literature ([22]).[1]

[1] In this article, whenever we simply refer to valency, we mean "active" valency.

After the first presentation of valency by Tesnière, the study of valency was taken up by many scholars, with a wealth of material now available. Since individual authors see valency from different perspectives, so far no generally accepted definition of valency exists (Storrer [36]). Generally, valency is understood as a specific ability of certain lexical units - primarily of verbs - to open free slots for filling in by other lexical units. By filling these slots a sentence structure is being built. Valency is seen as both syntactic, semantic, or some combination of them.

The valency terminology is also inconsistent; terms like valence, subcategorization, intention (in [30]), government, government pattern ([20]), complex sentence pattern ([4]), argument structure ([26]), stereotypical syntagmatic patterns ([32]) etc. emerge. Naturally, these terms not always denote exactly the same linguistic phenomenon. For a detailed survey see [40].

1.2 Valency in the Functional Generative Description

Among theories combining the syntactic and semantic approach, the valency theory developed within the framework of the Functional Generative Description (FGD) is found. (see e.g. [28], [29] and [16]). It uses syntactic as well as semantic criteria to identify verbal complementations. In this theory, it is assumed that potentially every (semantic) verb, noun, adjective and adverb (i.e. every complex node) has subcategorization requirements, expressed by its valency frame. Valency modifications include all kinds of elements (dependency relations) that can modify a particular lexical unit.

For example, in the sentence *Jitka mu daruje knihu* (lit. *Jitka gives him a book*) the verb *darovat* (lit. *to give*) opens a slot for a subject in nominative, i.e. for an agent (*Jitka*), then a slot for a dative object, i.e. for an addressee of giving (*mu*, lit. *him*), and lastly a slot for an accusative object, i. e. for an object which is being given (*knihu*, lit. *a book*).

Since the FGD does not work with the notion of "semantic roles" as known from some of the literature (such as a "runner", "giver", "object-given", see e.g. [15]), the appropriate lexical unit (here the verb) therefore determines—besides the morphological requirements on arguments—also their semantic properties.

In FGD, we work with TECTOGRAMMATICAL REPRESENTATION of sentences, which reflect their underlying syntax and certain types of semantic attributes. In this formalism, the central position in a sentence (or clause) is occupied by a (typically) finite verb.[2].

In order to sort out the behavior of all word modifications and in order to describe their character we define the following main basic principles in our valency concept:

- a valency frame is assigned to each verb sense separately,[3]

[2] Also some nouns, adjective and adverbs valency frames are recorded in the PDT-Vallex, but we don't discuss them in this contribution.

[3] Verb senses are defined rather coarsely, as opposed to some other approaches, such as the famous WordNet resource. However, it is not excluded that two clearly distinct senses carry identical valency frames. In other words, senses are not forced to be merged just because their valency frames are the same.

- criterion for distinguishing inner participants (arguments) and free modifications (adjuncts),
- criterion for distinguishing obligatory and optional modifications, and
- the concept of "argument shifting".

According to the type of dependency, any modification can be classified as either an INNER PARTICIPANT (that is, an argument) or as a FREE MODIFICATION (which is close to what is known as an adjunct). A given modification of a particular lexical unit may be—with respect to its particular governing word—either OBLIGATORY (that is, obligatorily present in the deep, tectogrammatical structure) or OPTIONAL (that is, not necessarily present). For the obligatory vs. optional distinction, the DIALOG TEST, described later, is used.

The distinction between inner participants and free modifications is *not* verb-specific: if a dependency type is an inner participant, then it is considered an inner participant for all verbs which it possibly modifies. We have determined that there are five such types of arguments: actor (ACT), patient(PAT), addressee (ADDR), effect (EFF), and origin (ORIG). These arguments have also the additional property that they can appear only once in a given clause headed by the verb (in the particular verb sense) to which they belong.

Among the 70 complementation types used in the Prague Dependency Treebank, we identify 36 verb free modification types (adjuncts): adjuncts expressing semantic time relations: TFHL, THL, THO , TFRWH, TOWH, TPAR, TSIN, TTILL a TWHEN; adjuncts for local semantic relations: DIR1, DIR2, DIR3 a LOC; adjuncts for causative relations: ACMP AIM, CAUS, CNCS, COND a INTT; adjuncts for means relations: CPR, CRIT, DIFF, EXT, MANN, MEANS, REG, RESL a RESTR; modal adjuncts: ATT, INTF a MOD; semantically different adjuncts: BEN, CONTRD, HER a SUBS, and finally, adjuncts with double semantic dependency (verb and another verb agrument): COMPL.

While arguments can modify just a relatively closed class of verbs, every adjunct can modify (in principle) any verb. That is also where their name (free modification) comes from; moreover, they can be repeated within a given clause several times.

For distinguishing among the five inner participants we use syntactic as well as semantic criteria. Actor (ACT) is always the first inner participant (something like Arg0 in the PropBank) and Patient (PAT) is always the second inner participant (usually like Arg1 in the PropBank, see [26]). These two arguments are thus determined more or less syntactically. Only when a verb has more than two arguments, semantic criteria come into play. Semantic origin (for example, *to make of wood*) gets the label ORIG, semantic addressee (*talk to somebody*) gets ADDR and semantic result (effect) gets the label EFF (*to split into pieces*).

To stress the distinction between the typology of the first two arguments of the verb and the rest (if any), FGD has adopted the concept of shifting of arguments. According to this special rule, semantic Effect, semantic Addressee and semantic Origin (which would normally be labeled by EFF, ADDR and ORIG, respectively) are being shifted to the Patient position in case the verb has only two arguments. In the sentence *Peter has dug a hole* the semantic Effect (*a hole*) happens to be labeled PAT (as it is in all sentences headed by the same sense of the verb *to dig*); similarly, in the sentence *The*

teacher asked the pupil, the semantic Addressee is shifted to the Patient position. This rule simply helps to keep consistency at the expense of lower semantic "precision".

Both arguments and adjuncts can be in their relation to a particular word either obligatory (i.e., obligatorily present at the tectogrammatical level of sentence representation) or optional (i.e. not necessarily present in each sentence where the verb is used). It should be stressed that this does not concern the surface appearance of such modifications, because they can be elided virtually anywhere; the notion of obligatoriness is used in the semantic sense. A natural question arises how the obligatoriness can then be determined, given that surface appearance cannot be used as a criterion: we rely on a DIALOG TEST [27]. The dialog test is a method based on a question about something that is supposed to be known to the speaker because it follows from the meaning of the verb the speaker has used. If the speaker can sensibly answer a hearer's follow-up question about a semantically obligatory modification "I don't know", then it means that the given modification is optional. On the contrary, if the answer "I don't know" is not possible in the particular point of this dialog-to-be, then the given modification is considered obligatory. For example, if the verb *to leave* (in the sense of "departing") is used in a sentence *John left*, the speaker must know from where John left (otherwise, he or she would—even should—have used another verb). Consequently, "from where" (DIR1) is an obligatory modification. Conversely, the speaker does not need to know to where John left—thus, if present, the "to where" (DIR3) modification will always be optional.

1.3 Valency and the Prague Dependency Treebank

The concept of the valency frames in the Prague Dependency Treebank (PDT) annotation ([21]) corresponds to the valency theory built in the FGD framework described above.

The work on the valency lexicon enabled the confrontation of the valency theory and real usage of language. Thus, we can say that PDT-Vallex has been created "bottom-up"; it was not necessary to make up valency complementation examples for the theoretically given schemes of valency frames because the lexicon draw upon the real texts from a real corpus.

Primarily, the PDT-Vallex served for keeping inter-annotator consistency high during the process of manual corpus annotation, most importantly for functor assignment to verbal complementations. After the tectogrammatical annotation process has ended, the lexicon served also for rigorous, automatic cross-checking of the annotated PDT data against this newly built lexicon.

The PDT-Vallex contains only those words (verbs, nouns, adjectives and adverbs) and their senses which occurred in the annotated data. The lexicon contains 10039 different words: 5510 verbs, 3727 nouns, and a small number of adjectives and adverbs. The total number of valency frames is 14979, out of which there are 8810 valency frames for verbs.

The valency modifications are described in the valency frame of the particular verb. Arguments (inner participants) are always recorded, be they obligatory or optional; adjuncts (free modifications) are recorded only if determined obligatory.

Apart from the obligatoriness indication we also record the dependency relations (the FUNCTOR) and the morphemic surface form. Every verb has at least one valency frame; each frame corresponds to one sense (meaning) of the verb.[4]

For example, the (English) verb *to leave*, which has two clearly distinguishable senses, would have two valency frames in our valency lexicon. The first one would be used for the sense *somebody left something* (with an Actor and a Patient as the two obligatory arguments) and the second one for the sense *somebody left from somewhere* (with and Actor and Direction-from as the obligatory arguments for this sense).

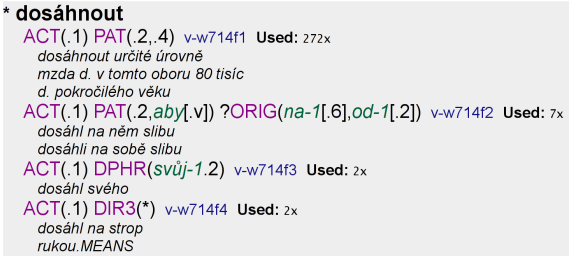


Fig. 1. The PDT-Vallex entry for *dosáhnout* (*to reach*)

A real PDT-Vallex entry for the verb *dosáhnout* (*to reach*) can be seen, formatted for better readability, in Figure 1. This verb has four different senses in our dictionary. The first sense *to reach something* corresponds to the first frame, which contains ACT in nominative and a PAT in genitive or in the accusative morphemic case. This frame has been used 272 times in the data. Below the formal description of the valency frame, three usage examples can be found. Similarly, the other three senses are described using the same structure.

An example sentence with the verb *dosáhnout* used as the main verb is in Figure 2. Obviously, this is an example of the usage of the most frequent sense (first entry as seen in Figure 1). The Actor (ACT) is the word *kurs* (*price*), and the other obligatory argument is the Patient (PAT)—the word *hodnota* (*value*), further modified by the actual price tag and the currency designation.

In Figure 3, we have schematically depicted how the corpus is linked to the PDT-Vallex lexicon. Let's say that in the corpus, we have 3 occurrences of the verb *uzavřít* (lit. *to close*) in 3 sentences. There are two PDT-Vallex entries (valency frames) for *uzavřít*. The first two occurrences are linked with the second valency frame with the basic meaning of *to close*, which has the usual transitive frame with two arguments: ACT and PAT. The third occurrence of *close* is linked with the first valency frame, which represents the light verb meaning, denoted here with the CPHR functor in its frame. The

[4] In rare cases, the description of which is beyond the scope of this article, two or more valency frames may still be used with the same sense of the verb. This is however not reflected in the current format of PDT-Vallex. For a suggestion of possible restructuring to allow for (i.a.) such grouping, see [42].

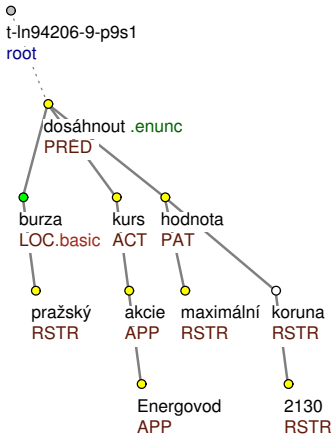


Fig. 2. The (simplified tectogrammatical representation of the) sentence *Na pražské burze dosáhl kurs akcií Energovodu maximální hodnoty 2130 korun.* (Lit. *On the Prague Stock Exchange reached the price of shares of Energovod the maximum value of 2130 Czech crowns*)

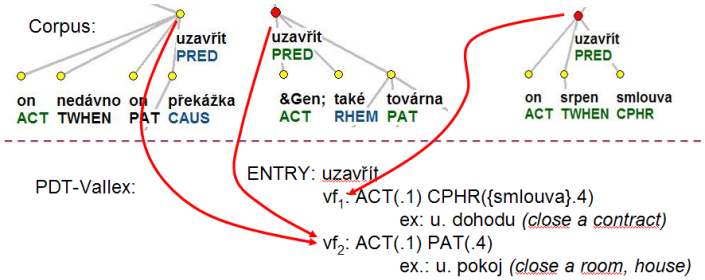


Fig. 3. Links between the corpus and the PDT-Vallex entries

arguments are linked implicitly, and their correctness in form can be determined using the valency entries and the diathesis transformation rules described in this article.

A detailed description of the surface form of a valency modification as captured in the valency frames in the PDT-Vallex can be found in Sect. 3.1 of this article, because this information is relevant for the description of diatheses and their surface transformation rules; for the most detailed information, see the annotation manual ([21]).

2 Diathesis

2.1 Diathesis in general

In contemporary linguistics, the very term DIATHESIS is closely related to other terms, e.g. alternation ([15]), conversion ([1]), hierarchization ([5]) or genus verbi ([12]). The

phenomenon of diathesis was elaborated in detail in the aforementioned work of F. Daneš and in other books and articles ([4], [35], [6], among others). Lately, this question is researched in relation to the valency lexicon [14]. Worldwide, the diathesis issue is elaborated e.g. by Meľčuk [19], Chrakovskij [3], Padučeva [24] and [23], Uspenskij [41] or Babby [2].

The DIATHESIS is understood as a syntactic grammatical category related to verb voice. It is defined, e.g., as the "relation among the elements of the semantic structure of the sentence and their corresponding syntactic positions" ([13], p. 522).

A given proposition can appear in the syntactic structure of the utterance either in its primary (basic, or canonical) diathesis and in a number of secondary (derived) diatheses. The primary diathesis is defined as the use of the verb in active voice (and in finite form). More general definition (12) says that the primary diathesis is the one which co-occurs with the highest number of complements on the surface, with the subject to the left of the verb.[5] Other diatheses, in which the semantic (deep) subject (or ACT in our case) is *not* in the subject position in the surface structure, are considered secondary. In Czech, the secondary diatheses are signaled by specific verb forms (such as the passive voice) and syntactic structures which force the semantic subject to move out of the surface subject position.

Specifically, periphrastic or reflexive passive constructions, constructions with the verb "to have" with verbal passive participle (resultative, similar to perfect tense in English), dispositional modality constructions, or the construction "to get" with a passive participle (causative) are all examples of secondary diatheses in Czech.

Sometimes, alternations are also classified as diatheses, such as in Kettnerová and Lopatková [14], where they distinguish grammatical diatheses (roughly, the ones mentioned in the previous paragraph) and semantic diatheses (alternations as the term is defined elsewhere).

The Czech valency dictionaries, both printed ([37], [17]) and electronic ([18], [25], [11]), contain (if ever) just canonical forms of the verb complementations used in the primary diathesis. The only exception is the dictionary of Skoumalová [34], who captures also the explicit complementation forms used in the secondary diathesis. However, for the natural language processing (as well as for verifying the theoretical language description) a valency lexicon with a systematical description of all syntactic and morphosyntactic forms is needed; the analysis and synthesis of Czech language, i.e. information about the morphematic realization of particular verb complementations, would be helpless without this piece of information.

3 Transformation rules for diathesis in PDT-Vallex

In this section, we first describe the means for formalizing form of expression of the verb and especially its arguments in PDT-Vallex. Then, after describing the general ideas behind formalizing form transformation occurring in diatheses, we describe the types of diatheses that we have dealt with in PDT-Vallex in more detail and give some examples.

[5] In Czech, this implies the standard word order, i.e. with the subject not being the focus of the sentence.

3.1 Explicit description of surface form in PDT-Vallex

As it has been said already, the PDT-Vallex entries describe, in a fully formalized way, the canonical (primary diathesis) expression of the verb and its arguments. However, with only a few extensions, the same formalism can be used for the explicit description of the necessary form of expression of secondary diatheses (i.e., the result of a transformation of the canonical surface form to the secondary diathesis one).

In all cases, the description of form should in general describe the predicate and its arguments as a whole, due to various possible interdependencies among the expression of form for the verb and its individual arguments ([8])) . However, with only a few exceptions, the form is independent among the arguments of the verb. Therefore we decided that the description of form will be associated with the individual arguments, independently of each other. This is true for both the canonical form (as present in PDT-Vallex) as well as for the resulting transformed form descriptions for the diatheses, even though the transformation rules themselves have to consider several or all the arguments at once.

The part-of-speech and morphosyntactic requirements which characterize the surface form corresponding to the valency slot (argument) in question are denoted by a short, formally defined string of symbols, separated from the m-lemma (if any) by a separator (mostly a period, `(.)`). They appear in the following order: part of speech requirement, then the morphosyntactic requirements (values) of gender, number, case, degree of comparison, agreement and negation. If any of these designators is missing, any value of the given category is allowed (in most cases, that means it is not really relevant for the relation between the verb and the argument in the corresponding slot).

The first designator (for part of speech) sometimes carries some additional requirement, such (for a verb) to be an infinitive. At the part-of-speech position, clausal restrictions (if the complement is realized as a clause, and not as a noun or other simple phrase). Lowercase letters are used at this position:

- `a` adjective
- `d` adverb
- `n` noun
- `i` interjection
- `v` verb
- `f` verb in infinitive
- `u` possessive pronoun or adjective
- `j` subordinate conjunction (with a clause it governs)
- `s` direct speech (root of subtree)
- `c` relative clause (root of subtree)

Gender is denoted by the following four capital letters: `F` for feminine, `M` for masculine animate, `I` for masculine inanimate and `N` for neuter.

Number is denoted by uppercase `S` and `P` with the obvious meanings.

For (morphosyntactic) case, digits are traditionally used for the seven cases in Czech: `1` for nominative, `2` for genitive, `3` for dative, `4` for accusative, `5` for vocative, `6` for locative and `7` for instrumental. The degree of comparison also uses the digits `1` to `3`

for positive, comparative and superlative, respectively, preceded by the symbol @ to distinguish them from the case markers.

Agreement in gender, number and case with the governing node at the surface layer is denoted by #.

For negation, we use the tilde character (~).

In addition, any combination of the morphological attribute values that are used at the morphological and analytical layers of the PDT can be included as a requirement. Special marking separates them from the above shorthands: a $ symbol and a tag index in < and > must precede the concrete value, which then should match directly the tag position at the given index.

The surface form designation might have alternatives (separated by a semicolon), or even be empty. Empty form designation is allowed only for free modifications, and it means, that any form that is associated with the particular functor can be used.

Examples of the designation of requirements on the surface realization, roughly sorted by frequency of appearance in the PDT Vallex dictionary:

1. Case-only requirement: .4
2. Preposition and a particular case: s[.7]
3. Alternative surface expression: preposition (with a particular case), or a case-only designation: pro[.4];.3
4. A particular subordinate conjunction (alternative of two) governing a verbal clause on the surface: že[.v];aby[.v]
5. Dependent clause, no conjunction: .v
6. Multiword preposition with genitive: od-1[na-1,rozdíl,.2]
7. Phraseme *balit fidlátka* (lit. *pack one's belongings*, i.e. *to leave*): fidlátko.P4

It should be also noted that in reality, the form designators described above are rather short abbreviations for sometimes much more complicated logical expressions; for example, .1 and .4 are matched also by various prepositional or nominal forms of numerical expressions not necessarily in nominative or accusative.

In addition, a preposition requiring a single case can be abbreaviated as <preposition>+<case> (e.g., as in od+2, corresponding to the less readable format od[.2]). For the purpose of brevity, we similarly introduce (in this paper only) a single number (for example, 4) to mean a non-prepostitional, direct object in the given case (normally written as .4).[6]

3.2 Diatheses and form transformation

Again, the PDT-Vallex dictionary [39], contains only the canonical surface form designation, i.e. the one which describes the form of the verb complements in the primary diathesis appearance (active voice, finite form). This seemingly causes inconsistency with the corpus annotation, since the form as required by the form designator in the valency frame pointed to by the occurrence of the verb sense in the corpus does not match the actual form of the complements in the corpus if any of the secondary diatheses is used. However, since the change of the form when the verb appears in a secondary

[6] This latter abbreviation is not used in the real data, however.

diathesis does not depend on the particular verb or verb sense, we can use quite general "transformation rules", which can convert the form designation present at the canonical valency to its "secondary" designation which then should display a perfect match with the annotated occurrence in the corpus.

This has allowed for a complete verb sense and valency annotation of every occurrence of every verb in the corpus, while still being able to check for the correct relation between the verbal frame in the PDT-Vallex dictionary and its surface realization in the corpus even if the verb is not used in the primary diathesis form.

The transformation rules were prepared with annotation consistency checking in mind. Therefore, they aim at such transformation of the valency frame (using also the annotation information from the corpus) to arrive at a simple set of checks to either confirm or reject whether the annotation of the verb and its dependents and its context is consistent with the requirements found in the valency frame.

Every transformation rule has two parts:

1. condition to be fulfilled at the node being checked, and
2. a set of rewriting rules.

Every rewriting rule has three parts:

1. the type of the rule (replacement, alternative)
2. assertions about the verb frame
3. specification of the necessary changes in the valency frame

While we will not fully dissect all the rules in the following sections describing the individual diatheses according to the above structure, we will aim to describe all the main aspects of the transformation. Full details can be found in [40].

3.3 Transformation rules for periphrastic passivization

Only transitive verbs (i.e., verbs with a slot marked PAT in the PDT-Vallex) can appear in the periphrastic passivization type of diathesis.

The verb itself must be in the form of passive participle, and the actor (ACT) is moved from the subject position to either an instrumental-case object position (corresponding roughly to the English prepositional phrase with the preposition *by*) or it is realized as a prepositional phrase with genitive preposition *od* (lit. *from*). Sometimes, both forms are allowed ((7) as well as (od+2)). This transformation of form always takes place, regardless of what other complement gets to the subject position.

The (surface) subject can either be missing (zero pronoun form), or in fact any of the arguments can get to that position:

PAT *The painter painted a picture*.PAT
→ *A picture was painted by a painter.*
ADDR *The injury slowed down the athlete*.ADDR
→ *The athlete was slowed down by an injury.*
EFF *The teacher has read a resume*.EFF *about him*
→ *A resume has been read about him by the teacher.*

Periphrastic diathesis, if used with a perfective verb, can be also considered a resultative diathesis, if it describes a completed event. However, the aim of the use has no reflection on the form changes that the arguments undergo in this diathesis, therefore we are not making this distinction here.

In periphrastic passivization, which is by far the most frequent of all diatheses, the following cases are covered by our rules:

1. PAT is moved to the subject position; in this case, we have to further look at the form of the PAT actually found in the data. The individual forms will be transformed as follows:
   - (4) → (1): nominative case for the "moved" subject
   - (f) → (f): infinitive stays as such; verb agreement: 3rd. pers. Sg. Neuter
   - (c) → (c): relative clauses do not change; 3rd. pers. Sg. Neuter
   - other form of PAT → DELETE (i.e., does not appear on the surface)
   - other arguments: forms kept as they are recorded in the canonical frame, except if EFF(jako+4) (lit. *as* with accusative) is present in the valency frame together with PAT(4) → PAT(1), then it is changed to EFF(jako+1).
2. ADDR is moved to the subject position
   - (4) → (1): nominative case for the "moved" subject; for other forms of expression in the canonical form, no change.
   - other arguments: forms kept as they are recorded in the canonical frame
3. EFF is moved to the subject position
   - (4) → (1): nominative case for the "moved" subject; for other forms of expression in the canonical form, no change.
   - other arguments: forms kept as they are recorded in the canonical frame (except when EFF(jako+4) is used, see above at PAT).

Example of transformation:

požádat (*to ask*)	ACT(1) ADDR(4) PAT(o+4,aby)
	→ ACT(7) ADDR(1) PAT(o+4,aby)
říci (*to say*)	ACT(1) ADDR(3) PAT(o+6) EFF(4,že)
	→ ACT(7) ADDR(3) PAT(o+6) EFF(1,že)
přijímat (*to hire*)	ACT(1) PAT(4) EFF(jako+4)
	→ ACT(7) PAT(1) EFF(jako+1)

Example application to a sentence:

> *Rektor požádal tajemníka o dokumentaci.* → *Tajemník byl požádán rektorem o dokumentaci.* (lit. *The rector asked the secretary for the documentation.* → *The secretary was asked by the rector for the documentation.*)

Here, the ADDR (*the secretary*) moves to the subject position and its form changes from the canonical accusative to nominative (*tajemíka* → *tajemník*), whereas the ACT (*the rector*) must be then expressed in the instrumental case (*rektor* → *rektorem*). The PAT (documentation) remains in the prepositional phrase form, using the preposition *o* (*for*) and the accusative case (o+4).

Univerzita přijímala tyto cizince jako překladatele. → Tito cizinci byli univerzitou přijímáni jako překladatelé. (lit. *The university hired these foreigners as translators. → These foreigners have been hired by the university as translators.*)

Both the words *cizinci* and *překladatelé* appear in the passive construction in the nominative case, following the ACT(1) PAT(4) EFF(jako+4) → ACT(7) PAT(1) EFF(jako+1) rule.

3.4 Transformation rules for reflexive passivization

In Czech, reflexive passivization adds the particle *se* (lit. *itself*, but it should be noted that this word has lost its proper meaning in this construction) to the verb. It can only be applied to verbs which do not use the same particle in active form (*reflexivum tantum*), or inherent reflexives, where the reflexive meaning is lost completely, such as *smát se*, lit. *to laugh*).

In the transformation of form for reflexive passivization, the subject position is taken by some other argument than ACT, similarly to the periphrastic passivization. However, the ACT never appears on the surface; it is structurally excluded by the syntactic rules of Czech. In other words, this diathesis can be used by the speaker only in the case he does not need to explicitly mention the ACT argument in his utterance, such as in the case when it is general ("everybody").

It is quite common that either PAT or ADDR, which would normally be moved to the subject position, are dropped[7], too. In such a case, a place (LOC, DIR1, ..) or time (TWHEN, TTILL, ..) expression must be present[8], or at least understood from the context. This fact, however, does not change the form of transformation rules.

We do not repeat the individual rules for the transformation of form for the PAT, ADDR, EFF and ORIG arguments, since they are the same as in the periphrastic passivization. However, the following rules must be applied in addition to the periphrastic passivization ones:

1. the particle *se* must be added to the verb (as a separate word), with its word order determined by Czech grammatical rules;
2. the phrase corresponding to ACT must be completely dropped on the surface (i.e., at the tectogrammatical representation the ACT must be represented by the #Gen t-lemma);
3. the tense of the verb remains active (or it remains in the infinitive, as the case may be);
4. the agreement rules on the surface are also determined by the Czech grammatical rules (e.g., if subject is not present at all on the surface, the verb must be in 3rd person singular form, and neuter gender if applicable).

[7] By "dropping" an ACT (PAT, EFF, ...), we mean that there is no word or phrase in the surface form of the sentence corresponding to the dropped argument. In the tectogrammatical representation of the sentence, a special t-lema and a special attribute are used to denote this fact.

[8] Unless the verb itself is in the proper focus of the sentence; for example, as a reply to the question *What do you do with books?*, one can say (only) *Books are read.* without adding when or where.

5. if PAT is moved to the subject position, the same rules apply as in the periphrastic passivization, except the ACT must be dropped on the surface.

Example application to a sentence:

Dělníci staví studnu z kamene. → *Studna se staví z kamene.* (lit. *The workers build the well from stone.* → *The well [itself] builds from stone.*)

The ACT (*Dělníci*) must be dropped completely, while the PAT (*Studna*) becomes the subject (in nominative, as usual).

Univerzita přijímala tyto cizince jako překladatele. → *Tito cizinci se přijímali jako překladatelé.* (lit. *The university hired these foreigners as translators.* → *These foreigners have [themselves] hired as translators.*)

The ACT (*Univerzita*) has been dropped in the reflexive passivization transformation. In addition and identically to the periphrastic passivization diathesis, both the words *cizinci* and *překladatelé* appear in the passive construction in the nominative case, following the ACT(1) PAT(4) EFF(jako+4) → ACT(7) PAT(1) EFF(jako+1) rule.

Děti o Vánocích zpívají koledy. → *Koledy se zpívají o Vánocích. Children at Christmastime sing carols.* → *Carols [themselves] sing at Christmastime.*)

While the PAT undergoes the usual transformation from accusative to nominative, a time (or location) adverbial is usually kept (or "added") for the sentence to sound natural in the reflexive passivization form (for an exception, see the footnote on the preceding page).

3.5 Transformation rules for resultative diathesis

The resultative diathesis (which is normally expressed by the verbs *mít* (lit. *to have*), *dostat* (lit. *to get*) with passive participle of the main verb) is used to move the addressee to the surface subject position (sometimes also to hide the causativity, or the agent, of the event). It can only be used with transitive verbs, where addressee must be present (either as a true addressee (ADDR) with patient (PAT) also present, or shifted to the patient position). Moreover, the *dostat*-type of diathesis can be used for only a limited class of verbs.

The following transformation rules apply (ANY means that the change applies to whatever the original form of the argument was):

1. ADDR is moved to the subject position (no other argument can be moved to this position):
   - (ANY) → (1): nominative case for the "moved" subject
2. PAT, if it appears on the surface, keeps its form, and if in accusative, forces an agreement in gender and case of the passive participle of the main verb, since this argument becomes the complement (Atv, AtvV)[9] on the surface. Moreover, if the

[9] These are the surface syntax functions. For the description of the formalization of surface syntax, which is outside the scope of this article, see e.g. [7].

gender is feminine in singular, the passive participle must use the special accusative form *-u* (the only verbal form which is thought to have a morphosyntactic features of case). PAT can also be deleted on the surface; the agreement forced on the passive participle form is then neuter singular.

3. ACT is moved to surface object position with the usual form transformation:
   - (1) → (7;od+2)

   In the resultative transformation, however, the (od+2) form is much more frequent than the instrumental case form.
4. Forms of other possible arguments are kept as they are.

Example of transformation (*mít*, lit. *to have*):

připravit (*to prepare*) ACT(1) ADDR(3) PAT(4)
→ ACT(od+2;7) ADDR(1) ?PAT(4)

Example application to a sentence:

Otec připravil dceři školní tašku. → *Dcera měla školní tašku připravenu od otce.* (lit. *The father prepared [for] daughter the schoolbag.* → *Daughter had the schoolbag prepared by the father.*)

Example of transformation (*dostat*, lit. *to get*):

přidat (*to add*) ACT(1) PAT(4;na+6) ADDR(3)
→ ACT(od+2;7) ADDR(1) ?PAT(4;na+6)

Example application to a sentence:

Ředitel přidal na platu jen střednímu managementu. → *Jen střední management dostal od ředitele přidáno na platu.* (lit. *The director raised in salary only to middle management.* → *Only middle management got from the director raised in salary.*)

3.6 Dispositional diathesis (dispositional modality)

This type of diathesis is used when the speaker expresses the "modality" (extent, in the form of and adverbial) of the relation between the verb and its actor, typically when the actor is an animate object (a human). The adverbial, which must be present on the surface form once the diathesis is applied, expresses often the degree of difficulty (or ease) with which the actor can perform the given action or keep themselves in a given state.

The following transformation rules apply (ANY means that the change applies to whatever the original form of the argument was):

1. ACT is either elided on the surface (represented as a general actor), or if present, its form is changed to the dative case:
   - (1) → (3): dative case for the actor; often expressed by a pronoun, with no restrictions on the short/long form of the pronoun (standard grammatical rules apply).

2. PAT becomes the surface subject, with accusative becoming nominative:
   - (4) → (1); other forms unchanged.
3. forms for other arguments are unchanged. This might result in two dative-form arguments of the verb, which is normally avoided, but it cannot be excluded on purely grammatical grounds. Often, though, the other arguments are elided on the surface.

Two other conditions must also be fulfilled:

1. An adverbial expressing the degree of difficulty must be present on the surface, or implicitly but clearly understood from the context;
2. the particle *se* must be added to the verb (as a separate word), with its word order determined by Czech syntactic rules. However, if the verb is inherently reflexive (reflexivum tantum), it keeps its single reflexive particle *se* without adding another.

Example of transformation:

studovat (*to study*) ACT(1) PAT(4)
→ (se) ACT(3) PAT(1) adverbial-func(*),

where adverbial-func is typically a modification of manner (MANN).

Example application to a sentence:

Pavel studuje angličtinu. → *Angličtina se Pavlovi studuje snadno.* (lit. *Paul studies English.* → *English (itself) to-Paul studies easily.*)

3.7 Reciprocity

Strictly speaking, reciprocity is not a diathesis in the proper sense, since the changes on the surface are not related to the change in voice or other inflectional feature of the verb. It is considered a diathesis because it shares some of the features of the other types of diatheses (e.g. adding a particle *se* to the verb).

In reciprocity, a single entity (or a group of entities, syntactically expressed as a single phrase, perhaps a coordinated one) occupies two arguments of the verb at the same time: most often it is the actor (ACT) and patient (PAT), but other pairs are also possible, e.g. PAT and ADDR.

Except for adding the particle *se* (unless, similarly to the dispositional modality type, already inherently present), no other changes in form are visible, except that one argument (typically, the "later" one of the two arguments being involved) is missing in the surface and the other one must be either

- in semantic plural (i.e., in surface morphosyntactic plural or it must be a mass noun, which keeps its surface singular inflection);
- a coordination, typically a conjunction, unless all members of the coordination fulfill on of the conditions on this list;
- construction of a single noun phrase modified by another prepositional phrase using the preposition *s(e)* (lit. *with*); this is often considered just another form of coordination;

– a single noun phrase describing a group of people of objects.

In the sentence representation, the "missing" argument is represented by an artificial t-lema `#Rcp` and the appropriate functor (PAT, ADDR, etc.), so that obligatory arguments are present as required. The double-function argument keeps the first functor (ACT in case of ACT–PAT reciprocity, PAT in case of PAT–ADDR reciprocity, etc.) and its form has to fulfill one of the above alternatives. Often, but not necessarily, the adverbials *(sebe/sobě) vzájemně/navzájem* (lit. *(themselves) mutually*), *jeden druhého/druhému/...* (lit. *(to) each other*) or similar are used on the surface to explicitly stress the reciprocity.

Additionally, the forms of location adverbials are slightly changed to reflect the symmetricality of the event, but the description of such changes in free modifiers is beyond the scope of this article (and has not been dealt with systematically in the PDT either).

Example of a transformation:

vidět (*to see*) ACT(1) PAT(4)
→ (se) ACT(<sem-plural>1) PAT(-)

where <sem-plural> is the surface form of all possible semantic plurals, and (-) means there should not be any form corresponding to this argument on the surface (moreover, the t-lema should be `#Rcp`).

Example application to a sentence:

> *Pavel viděl Janu na druhé straně ulice. → Pavel a Jana se viděli přes ulici.* (lit. *Paul saw Jane at the other side of the street. → Paul and Jane saw (themselves) across the street.*)

4 Future work

In this arcticle, we have described valency in the PDT-Vallex valency dictionary, concentrating on its formal properties and especially on the change of surface form of the verb and its arguments when secondary diatheses are used. The report presented here is not exhaustive; more details can be found in [40].

In the future, we would like to fully formalize the transformation changes not only for the purposes of checking the consistency of annotation (as it was done in the Prague Dependency Treebank and its PDT-Vallex dictionary), but also to serve, in a simpler manner than today, the task of natural language generation, where diathesis and specifically its surface from are of high importance ([31]).

Obviously, we will be extending the PDT-Vallex dictionary further. For example, the dictionary is now being extended to cover all the verbs in the Czech translation of the Wall Street Journal portion of the Penn Treebank, which is part of a parallel treebank project called the Prague Czech-English Dependency Treebank.[10]. These extensions will also necessarily lead to more examples of diathesis, and thus might need extensions of the surface form transformation rules described in this article.

[10] `http://ufal.mff.cuni.cz/pcedt`

Acknowledgements

We gratefully acknowledge the support of the Czech Ministry of Education through the grant No. MSM-0021620838, the Grant Agency of Charles University in Prague through the grant No. GAUK 52408/2008 and the Academy of Sciences of the Czech Republic through the grant No. 1ET101120503.

References

[1] Apresjan, J. D. (1995). *Leksičeskaja semantika. Sinonimičeskie sredstva jazyka*. Moskva.

[2] Babby, L. H. (1998). Voice and Diathesis in Slavic. In *Position paper presented at the Workshop on Comparative Slavic Morphosyntax: State of the Art. Indiana University, Spencer*.

[3] Chrakovskij, V. S. (1981). Diateza i referentnosť. (K voprosu o sootnozenii aktivnych, passivnych, refleksivnych i reciproknych konstrukcij). *Zalogovye konstrukcii v raznostrukturnych jazykach*, pages 5–38.

[4] Daneš, F. (1971). Větné členy obligatorní, potenciální a fakultativní (Obligatory, Potential and Optional Constituents of the Sentence). *Miscellanea Linguistica. Acta Universitas Palackiana Olomucensis*, pages 131–138.

[5] Daneš, F. and Hlavsa, Z. (1987). *Větné vzorce v češtině*. Academia, Praha.

[6] Grepl, M. and Karlík, P. (1983). *Gramatické prostředky hierarchizace syntaktické struktury věty* . Brno.

[7] Hajič, J. (1998). *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank*. Karolinum, Charles University Press, Prague.

[8] Hajič, J. and Honetschläger, V. (2003). Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation. *Prague Bulletin of Mathematical Linguistics (PBML)*, 79–80:61–86.

[9] Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and Ševčíková-Razímová, M. (2006). *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia, PA, USA.

[10] Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, J. and Hinrichs, E., editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

[11] Hlaváčková, D., Horák, A., and Kadlec, V. (2006). Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. In *Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2006*, volume 4188, pages 79–86, Berlin and Heidelberg. Springer.

[12] Karlík, P., Nekula, M., and Pleskalová, J. (2002). *Encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha.

[13] Karlík, P., Nekula, M., and Rusínová, Z. (2000). *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny, Praha.

[14] Kettnerová, V. and Lopatková, M. (2009). Changes in Valency Structures of Verbs: Grammar vs. Lexicon. In *This Volume, Proceedings of Slovko 2009*. Springer-Verlag Berlin Heidelberg.

[15] Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

[16] Lopatková, M. and Panevová, J. (2005). Valence vybraných sloves pohybu v češtině. In Piper, P., editor, *Proceedings of Matica Srpska za slavistiky*, pages 1–8, Novi Sad, Serbia and Montenegro, Oct. 27-29.

[17] Lopatková, M., Žabokrtský, Z., Benešová, V., Skwarska, K., Bejček, E., Chvátalová, K., Nová, M., and Tichý, M. (2007a). *Valenční slovník českých sloves*. Karolinum, Praha.

[18] Lopatková, M., Žabokrtský, Z., Benešová, V., Skwarska, K., Bejček, E., Chvátalová, K., Nová, M., and Tichý, M. (2007b). *VALLEX 2.5 - Valency Lexicon of Czech Verbs, version 2.5*.

[19] Mel'čuk, I. A. (1987). *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.

[20] Mel'čuk, I. A. (2003). Actants. In *Meaning-text theory 2003. Proceedings of the First International Conference on Meaning-text theory*, pages 111–127.

[21] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., štěpánek, J., Urešová, Z., Veselá, K., žabokrtský, Z., and Kučová, L. (2005). Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical Report TR-2005-28, Univerzita Karlova v Praze, MFF, ÚFAL, Prague.

[22] Nasr, A. and Rambow, O. (2004). SuperTagging and Full Parsing . In *Proceedings of Seventh International Workshop on Tree Adjoining Grammar and Related Formalisms* .

[23] Padučeva, E. V. (2000). Verbs implying semantic role of result: correlation between diathesis and aspectual meaning. *Linguistische Arbeitsberichte 75*, pages 125–136.

[24] Padučeva, E. V. (2002). Diateza i diatetičeskij sdvig (diathesis and diathesis shift). *Russian Linguistics*, 26:179–215.

[25] Pala, K. and Ševeček, P. (1997). *Valence českých sloves*, pages 41–54. Brno.

[26] Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

[27] Panevová, J. (1975). On verbal frames in functional generative description II. *Prague Bulletin of Mathematical Linguistics*, (23):17–52.

[28] Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Academia, Praha.

[29] Panevová, J. (2002). Sloveso: centrum věty; valence: centrální pojem syntaxe. In neuvedeno, editor, *Aktuálne otázky slovenskej syntaxe*, pages x1–x5.

[30] Pauliny, E. (1943). *Štruktúra slovenského slovesa*. SAVU, Bratislava.

[31] Ptáček, J. and Žabokrtský, Z. (2006). Synthesis of Czech Sentences from Tectogrammatical Trees. In *Lecture Notes in Computer Science, Proceedings of the 9th International Conference, TSD 2006*, number 4188 in Lecture Notes in Computer Science, pages 221–228, Berlin / Heidelberg. Springer-Verlag Berlin Heidelberg.

[32] Pustejovsky, J. (1996). *Generative Lexicon*. MIT Press, Cambridge, Massachussetts.

[33] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel, and Prague, Academia, Prague.

[34] Skoumalová, H. (2001). *Czech Syntactic Lexicon. Ph.D. Thesis*. PhD thesis, Faculty of Philosophy, Prague.

[35] Štícha, F. (1984). *Utváření a hierarchizace struktury větného znaku*. Univerzita Karlova, Praha.

[36] Storrer, A. (1992). Verbvalenz. theoretische und methodische grundlagen ihrer beschreibung in grammatikographie und lexikographie. *Reihe Germanistische Linguistik*, 126:414.

[37] Svozilová, N., Prouzová, H., and Jirsová, A. (1997). *Slovesa pro praxi. Valenční slovník nejčastějších českých sloves*. Academia, Praha.

[38] Tesnière, L. (1959). *Eléments de syntaxe structurale*. Klincksieck, Paris.

[39] Urešová, Z. (2005). *Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View*, pages 93–112. Veda Bratislava, Slovakia.

[40] Urešová, Z. (in prep.). *Valence sloves v Pražském závislostním korpusu*. PhD thesis, Univerzita Karlova v Praze, MFF, Praha.

[41] Uspenskij, V. A. (1977). K ponjatiju diatezy. *Problemy lingvističeskoj tipologii i struktury jazyka*, pages 65–84.

[42] Žabokrtský, Z. (2005). *Valency lexicon of Czech verbs (PhD thesis)*. PhD thesis, Univerzita Karlova v Praze, MFF, ÚFAL, Prague.

A Corpus of Spoken Language and Its Usefulness in the Research on Language Contact

Marcin Zabawa

Institute of English, University of Silesia, Poland

Abstract. The aim of the present paper will be to discuss two seemingly distant phenomena, namely (1) spoken language corpora and (2) language contact, or, to be more precise, the research on the borrowings of English origin in spoken informal Polish. It would seem that the most objective method of studying loanwords in spoken language is the use of the corpus, as it enables a linguist to formulate hypothesis on solid bases. It is then not only possible to state the existence of a given feature, but also to provide the evidence that would not be available without a corpus. Unfortunately, however, the corpora of spontaneous spoken Polish are practically non-existent; as a consequence, the only solution that seems to be left is to create such a corpus on one's own. In my paper I would like to concentrate on discussing the process of collecting and analyzing a corpus of spoken language which can then be used for the research on English borrowings. Special emphasis will be placed on various traps and problems associated with the task in question, such as the choice of material, the choice of informants, the method of transcribing recorded texts, the (non-)representativeness of the corpus, etc. The present paper is based on the author's doctoral dissertation [18].

1 General information about corpus studies

The term 'corpus' (form Latin 'body') is defined differently by various specialists working in the field. Some of the definitions are very broad and general, e.g. 'any collection of more than one text' [10], cited also in [11], whereas others are narrow, but at the same time fairly comprehensive, e.g.

> a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language [1].

McEnery and Wilson [10] add that although a corpus can be understood as any body of text, the term is normally used in modern linguistics in somewhat restricted sense. In other words, a corpus must have some specific features, described under four main headings: sampling and representativeness, finite size, machine-readable form and a standard reference (for details, cf. [10]).

Corpora can be basically divided into three main groups: (1) containing only written texts, (2) containing only spoken texts (transcriptions) and (3) composed of both written and spoken language. It seems fairly obvious that it is the spoken language corpus that is particularly difficult to construct and analyze. According to Sinclair:

> Most corpora keep well away from the problems of spoken language – with some honourable exceptions – and, for a corpus which in any way purports to reflect a 'state of the language', this is most unfortunate. Many language scholars and teachers believe that the spoken form of the language is a better guide to the fundamental organization of the language than the written form; and many writers comment on the differences. In my own experience, there is no substitute for impromptu speech […] [17].

Nowadays large corpora, stored in computer memory, are an invaluable aid in many branches of linguistics, most notably in lexicography, as the majority of the new dictionaries (both mono- and bilingual) are based on corpora, e.g. *Collins Cobuild English Dictionary for Advanced Learners*, 3rd edition, ed. by Sinclair (2001), based on *the Bank of English*, a corpus consisting of around 400 million words or *Oxford Dictionary of English*, 2nd edition, ed. by Soanes and Stevenson (2003), based on *the Oxford English Corpus*, composed of *the British National Corpus* (100 million words), *the Oxford Reading Programme* (around 77 million words) and other databases.

Apart from lexicography, corpora can be successfully used in language teaching, historical linguistics, pragmatics and discourse analysis, sociolinguistics, stylistics, and many other spheres (for details, cf. [10]). They can also be used in the studies connected with language contact.

2 Corpora and language contact

The use of a corpus enables linguists to formulate hypotheses about language on solid bases. It is thus possible not only to state the existence of a given feature, but also to provide the evidence that would not be available without a corpus. As for the study of language contact (manifested through the existence of borrowings), a linguist working with a corpus typed into a computer is not only able to state the existence of a given loan in a language, but also to discuss its frequency, both absolute and relative (i.e. compared to other loans), usually in terms of types and tokens, conduct the contextual analysis, determine all the senses in which a word was used and rank them according to their frequency, discuss the relation between the usage of a loan and various other criteria, such as the types of texts, the topics, the informants (their sex, education, age, knowledge of foreign languages). Moreover, as Otwinowska-Kasztelanic [12] rightly notices, it is not only the existence of a given feature that may be linguistically and/or statistically significant, but the non-existence of it may also constitute an important observation.

As for the language contact, it can be stated that the influence of English upon Polish has been researched relatively thoroughly in Polish linguistics. Sadly, however, most of the studies concentrate on discussing individual instances of borrowings in Polish (usually based on written sources, e.g. the language of the press) and thus they do not constitute a systematic research on the subject. The situation is even worse in the case of the research of English borrowings in spoken Polish, as there are very few

studies on the subject. It would seem that the most objective method of studying loanwords in spoken language is the use of the corpus, as it enables a linguist to formulate hypothesis on solid bases. However, the corpora of spoken Polish are scarce, to say the least. This is the reason why it is in fact necessary to collect such a corpus on one's own.

3 Corpora of spoken Polish

As Dunaj notes (cited in [12]), the traditional linguistic studies in Poland were based solely on written Polish. As a consequence, there was hardly any research on spoken Polish before the 1960s (or even early 1970s), with the exception of dialectological studies. A similar view was expressed by Żydek-Bednarczuk [19], who states that the studies on spoken language were intensified only during the last 25 years, dealing mainly with phonetic realization, vocabulary, syntax and semantics.

At first, the studies on spoken language were connected with the research on dialects. It was only in the 1970s that the informal variety of Polish started to be considered as worthy of serious linguistic investigation. One could mention here such works as e.g. the study on the syntax of telephone conversation [14], or the research on text structure of informal conversation [19].

Before extensive recordings of spontaneous spoken Polish were carried out, some theoretical preliminaries were discussed, such as the typology of spoken Polish varieties (Buttler, Wilkoń, both cited in [19]), various types of language contact and language varieties ([6], [7]; cf. also [12]) or components of speech act in a text (Pisarkowa, cited in [19]). Following those introductory works, studies on urban Polish were carried out, notably in the following Polish cities: Katowice, Kraków, Łódź, Poznań and Warszawa (cf. [19], [12]). As a consequence, the following corpora of spoken Polish, among others, were published:

- in the region of Katowice: *Teksty języka mówionego mieszkańców miast Górnego Śląska i Zagłębia* [7], [8]
- in Kraków: *Wybór tekstów języka mówionego mieszkańców Krakowa* [2]
- in the region of Łódź: *Wybór tekstów języka mówionego mieszkańców Łodzi i regionu łódzkiego. Generacja najstarsza* [4], *Wybór tekstów języka mówionego mieszkańców Łodzi. Generacja starsza, średnia i najmłodsza* [5]
- in the region of Warszawa: *Korpus języka mówionego młodego pokolenia Polaków* [13].

However, there is still a shortage of corpora of spoken Polish, as the existing ones are relatively small, e.g. the corpus of informal spoken Polish of the Warsaw variety [13]. Naturally, there exist larger corpora of English, used primarily in lexicography, such as *Korpus języka polskiego* (`http://korpus.pwn.pl/`) or *Korpus IPI PAN* (`http://korpus.pl/index.php?page=welcome`). The problem is, however, that they cannot be treated as corpora of spoken Polish. In the case of the former

one, only 4.5% of words come from recorded conversations whereas the latter one does not contain spoken texts at all.

Consequently, as the corpora of spontaneous spoken Polish are practically non-existent (with the exception of the small corpora listed above), it would seem that – when one wants to analyze the borrowings used in spoken informal Polish – the only solution is to collect such a corpus on one's own. However, certain problems may (and will) arise. These are discussed in the next section.

4 Problems connected with spoken language corpora

When one decides to construct a corpus of spoken language and analyze it in terms of borrowed words and/or meanings, three difficulties are predominant: first, it is much more difficult to construct a representative corpus of spoken language, particularly consisting of informal conversations, than, say, of the language of the press. On one hand, to ensure a good quality of the recording and make it in accordance with the law, it should be carried out openly, i.e. non-surreptitiously. On the other, the presence of a tape recorder may and often does have influence on the linguistic behaviour of speakers, so it is necessary to make sure that the speakers do not feel inhibited and behave as naturally as possible. Moreover, it is often not easy to find appropriate people that would act as informants. They should be varied as to their sex, age, education, place of work, social and economic background and so on. Second, the creation of a corpus of spoken language is a difficult and time-consuming process, as the conversations must be recorded, transcribed and finally typed into a computer. The last two phases are particularly time-consuming, especially when compared to the process of creating a corpus of written language, which does not need to be transcribed and, instead of keyboarding, one can use faster and more efficient methods, such as optical scanning, i.e. machine reading (suitable especially for printed books, as Sinclair [17] notices) or the re-use of the material already in electronic form (suitable especially for press texts, which can frequently be found on the Internet). One could of course make use of such quasi-spoken varieties as interviews or film scripts but, as Sinclair [17] notes, they are not really instances of spoken, but rather written-to-be-spoken language; hence they do not have typical features of spontaneous informal conversations and could not be included in the corpus of spontaneous language:

> If it is impossible in an early stage of a project to collect the spoken language, then there is a temptation to collect film scripts, drama texts, etc., as if they would in some way make up for this deficiency. They have a very limited value in a general corpus, because they are 'considered' language, written to simulate speech in artificial settings. Each has its own distinctive features, but none truly reflects natural conversation, which for many people is the quintessence of the spoken language. There is special integrity in a text which is a full record of a public meeting, enquiry, court case, radio or television station, etc., despite the mix of impromptu and considered language that is used –

> scripts and even read-out statements are common. But such records are not likely to be representative of the general usage of conversation. [17]

The above quotation answers the question as to why it was decided to exclude such texts as interviews or discussions emitted e.g. by the television from the corpus compiled by the present author.

Third, a typical spoken-language corpus does not contain many borrowed words, as Poplack et al. pointed out:

> Gathering enough data for the systematic study of the use of borrowed words in a speech community is inherently very difficult. Tokens of these words are typically rare in monolingual discourse, so that several hours of speech will yield only a few dozen, most of which occur only once. In certain contexts, of course, and for certain topics of conversation, there will be some set of borrowings which are used repeatedly, but the imposition of contextual or topical restrictions would vitiate the comprehensiveness and representativity of any investigation attempting to give a general characterization of borrowing and integration. [16]

A similar view was expressed by Dunaj:

> Każdy, kto zajmował się badaniami języka mówionego, wie, że uzyskanie obfitych materiałów leksykalnych z wypowiedzi mówionych nie jest łatwe. Wymaga żmudnych, długotrwałych obserwacji. [3]
>
> Any person conducting research on spoken language knows that it is not easy to obtain rich lexical material from spoken utterances. It requires long and laborious observations. [my translation]

The above quotations answer the question as to why corpora of spoken language are infrequently used for the systematic study of loans in a language. Instead, as Poplack et al. [16] notes, most studies dealing with the use of borrowed words in spoken language resort to three alternatives:

(1) Artificial methods of the elicitation of data, used by e.g. Poplack and Sankoff [15], who prepared a random series of photographs of 45 everyday items which could be designated by concrete nouns. The bilingual (Spanish-English) informants were then asked to name the object and provide any additional words for the same concept they could think of.

(2) The analysis of a few isolated borrowings that occurred naturally, a method used by e.g. Mougeon et al., cited in [16].

(3) The analysis of anecdotal lists of borrowed words, a method used by e.g. Nash, cited in [16].

Nevertheless, it seems that such methods as e.g. using photographs to elicit nouns from the respondents can provide only indirect data and thus the results obtained in this way are not necessarily representative of the use of loanwords in informal spontaneous conversations. As a consequence, it was decided for the purpose of my doctoral dissertation to gather a corpus of spontaneous conversations, since there is still a shortage of studies dealing with the use of borrowings, particularly semantic ones, in spoken discourse. Some of the recordings included in the corpus (6 out of 20)

are connected with the topic of computers, the Internet or modern technology so as to ensure that a larger number of lexical and semantic loans will be available for analysis (cf. also Section 5.3 dealing with the topics of the recordings).

5 The corpus compiled by the author

5.1 Introduction

To ensure the reliability and homogeneity of the corpus, it was decided that only natural spontaneous conversations will be included. This automatically excludes not only all written-to-be-spoken forms (such as film dialogues) but also conversations conducted in a formal setting, such as interviews, discussions or talk shows emitted by the television.

In general, the corpus is similar to the use collected by Otwinowska-Kasztelanic [13]. Her corpus, however, consists of conversations conducted in Warsaw whereas the one compiled by the present author consists of the ones recorded in Upper Silesia.

The entire corpus consists of twenty recordings (60,564 running words altogether). One could possibly state that the size of the corpus is too small to draw general conclusions concerning the use of lexical and semantic loans in spoken Polish. It seems, however, that it is large enough to highlight certain tendencies. Besides, the entire corpus was collected, transcribed and analyzed by only one person, namely the author of the present paper. As a consequence, it could not be very large for practical reasons.

5.2 The informants

Altogether, 48 people (including the author of the study) participated in the conversations (31 women and 17 men). They are uniquely coded throughout all the recordings. In other words, the same person is given the same symbol in all the conversations in which he or she took part. The codes for female speakers are F1, F2, …, F30, F31 whereas the codes for male ones are M1, M2, …, M16, M17.

It must be noted at this point that nine speakers were not taken into account in the present study. They were accidental speakers (e.g. people asking for something), who uttered only a few words altogether. The number of speakers taken into consideration is thus 39 (25 women and 14 men). The basic information about the speakers (gathered in 2003) is presented below. The following information is given: the symbol of the respondent, his/her age, sex, level of education (and the field of education in brackets) and his/her knowledge of English. The knowledge of English of most of the informants was established by means of informal conversations and/or short placement tests. With this end in view, a four-point scale was used: none, basic, intermediate and advanced. It must be added that most of the speakers from the first group know a few English words and expressions, at least passively. The last group, in turn, includes also informants fluent in English.

M1	26	male	higher (physical education)	intermediate
M2	36	male	higher (history)	basic
M3	25	male	higher (English studies)	advanced
M4	25	male	higher (physical education)	basic
M5	not taken into account in the present study			
M6	26	male	higher (English studies)	advanced
M7	23	male	higher (computer science)	intermediate
M8	54	male	higher (law)	none
M9	38	male	secondary (technical education)	none
M10	not taken into account in the present study			
M11	49	male	secondary (technical education)	basic
M12	not taken into account in the present study			
M13	24	male	secondary (technical education)	none
M14	27	male	higher (computer science)	basic
M15	26	male	higher (physical education)	intermediate
M16	50	male	secondary (technical education)	none
M17	22	male	secondary (general education)	basic
F1	26	fem.	higher (psychology)	intermediate
F2	40	fem.	higher (Russian studies, English studies)	advanced
F3	25	fem.	higher (biology)	basic
F4	27	fem.	higher (English studies)	advanced
F5	55	fem.	higher (biology)	none
F6	46	fem.	higher (history)	none
F7	29	fem.	higher (psychology, English studies)	advanced
F8	not taken into account in the present study			
F9	25	fem.	higher (biology)	intermediate
F10	29	fem.	higher (Polish studies)	intermediate
F11	31	fem.	higher (theology)	basic
F12	29	fem.	higher (Polish studies)	intermediate
F13	35	fem.	higher (chemistry)	none
F14	not taken into account in the present study			
F15	36	fem.	higher (library science)	basic
F16	27	fem.	higher (German studies)	basic
F17	48	fem.	higher (German studies)	basic
F18	not taken into account in the present study			
F19	not taken into account in the present study			
F20	not taken into account in the present study			
F21	69	fem.	primary (general education)	none
F22	36	fem.	vocational (gastronomy)	none
F23	not taken into account in the present study			
F24	47	fem.	vocational (technical education)	none
F25	51	fem.	higher (medicine)	basic
F26	23	fem.	vocational (hairdressing)	basic
F27	23	fem.	vocational (gastronomy)	basic
F28	24	fem.	vocational (hairdressing)	none
F29	35	fem.	higher (geography)	basic
F30	50	fem.	higher (law, English studies)	advanced
F31	24	fem.	higher (marketing and management)	intermediate

As one can see, most of the informants are university graduates, but there is also a small percentage of people having secondary or vocational education. The age of most of the speakers (27 out of 39) ranges between 21 and 36.

To make the corpus conversations as varied as possible, it was decided to choose respondents representing various occupations, including the following trades and professions: primary and secondary school teachers (of various subjects), a university teacher, a computer programmer, computer specialists, office workers, an entrepreneur, an interpreter, a lawyer, a bank clerk, security guards, an electrician, a lorry driver, an electronics engineer, shop assistants, a psychologist, a cleaner, a cook, a doctor (physician), hairdressers, an old-age pensioner and unemployed persons. All of the speakers come from Upper Silesia in Poland.

5.3 The recordings

Altogether, there were twenty conversations recorded for the purpose of the study. Most of the conversations were recorded in informal situations, happening both indoors, during such events as a birthday party, a family meeting or a meeting with friends, cf. Recordings 5, 12 and 20 (12,022 words in total) and outdoors, during walks, often with a dog, e.g. in a park or the woods, along the lake shore, through a housing estate etc., or during informal meetings with friends and/or family members e.g. in a garden, cf. Recordings 1, 6, 9, 14, 16, 18 and 19 (25,871 words in total). Some other recordings, viz. number 2, 10 and 13 (6,581 words in total), were conducted in a place of work, namely in a staffroom in a school during long breaks. The situation was thus more formal than in the previous cases, but nevertheless the conversations could be safely described as informal, as most of the teachers taking part in them were on first name terms with one another. The rest of the recordings, namely number 3, 4, 7, 8, 11, 15 and 17 (16,090 words in total) were conducted between the students learning German in a foreign language school. The conversations were spontaneous and informal (recorded before the actual classes), as all the people involved were again on first name terms.

The topics of particular recordings and the number of running words and are presented below. The sequence of information is as follows: the number of the recording, the topic or topics of the conversation, the number of running words in a given recording.

1	dogs, holidays, sport contests, fishing	3966
2	school, teachers, pupils	2868
3	computers, the Internet, computer programs	1620
4	student exchange, computers and e-mail, exams and cheating	2027
5	acquaintances, family matters, working, excursions, holidays	3290
6	computers, computer games, computer magazines, the Internet	3654
7	computers and e-mail, student exchange, learning German	2348
8	teaching and learning foreign languages	3144
9	a wedding, doing shopping, business matters	3327

10	school, pupils and teaching, doctors and health-care	2178
11	correspondence, computers (esp. using a text processor)	2393
12	family matters, acquaintances, keeping dogs at home	3848
13	films, teaching and giving grades, marriages	1535
14	looking for work, doing business, computers and the Internet	4060
15	weather, holidays	2229
16	looking for work, learning abroad, moving to another country	3895
17	learning foreign languages, taking exams	2329
18	birthdays, films, new technologies, software piracy	3807
19	grilling and smoking food, working, ticks and other insects	3162
20	everyday life, cooking and eating, family, taking care of a baby	4884

As one can see, it was decided to include conversations on various topics. Three of them, however, seem to be prevailing: (1) everyday activities, such as working, living, talking about one's family, friends and/or acquaintances (Recordings number 5, 9, 12, 19, 20), (2) computers and the Internet (Recordings number 3, 4, 6, 7, 11, 18) and (3) teaching and learning (Recordings number 2, 8, 10, 13, 17). The conversations about computers have turned out to be particularly interesting for the study, as they contain a large percentage of both lexical and semantic loans found in the corpus (compared to other recordings).

5.4 The type of language used

As was noted above, the entire corpus has been composed of spontaneous conversations, hence the language used in it can be safely described as naturally occurring Polish. It should also be underlined that the language of some of the speakers has visible features of the Silesian dialect (notably the one of F21, F22, F24, F25, F28, M6, M7, M8, M9, M11, M13 and M16). It must be added that the situation of most of the speakers taking part in the study can be characterized as diglossic; in other words, the informants are able to speak both standard Polish and the Silesian dialect and alternate between them according to the situation. Some recordings are thus examples of naturally occurring Silesian dialect (especially number 1, 5, 6, 9, 12, 14, 18 and 19).

For the sake of clarity, only some most noticeable features of the Silesian dialect are marked in the corpus, e.g. some aspects of the special pronunciation (such as special endings or extra consonants), some untypical aspects of morphology or the use of specific vocabulary.

5.5 The transcription and the use of fonts

As the aim of the research was to study lexical and semantic loans in informal spoken Polish, it was not necessary to transcribe the recorded conversations phonetically. Instead, it was decided to use orthographic transcription. However, no punctuation marks were used, with the exception of a question mark, used for marking questions (indicated by the intonation used by a speaker). Moreover, the slash symbol (/) was

used to show short (up to one second) pauses in the speech. Such a notation is aimed at reflecting the natural flow of speech. The convention used for the transcription of the recordings used for the purpose of the present study is thus very similar to the one used by Żydek-Bednarczuk [19], who conducted a study on the structure of informal conversation.

A speaker's turn is always indicated by his or her code, i.e. a letter (F for women, M for men) followed by a number and a colon, all written in bold for easy reference. All names, surnames, nicknames and some other expressions denoting people (with the exception of some of them referring to celebrities) are replaced by common nouns given in curly brackets (e.g. {name}, {surname}, {nickname}, {company name}, {city}) in order to keep the informants and the people they speak about anonymous.

Other features marked in the transcriptions include: the pronunciation of items borrowed from English, given in angle brackets; non-verbal behaviour of the speakers, such as laughter, a cough, a whistle, etc., with the indication of its duration (in general, the duration of the phenomena lasting up to one second is marked as '[1 sec.]') given in square brackets; semi-verbal behaviour, expressing some kind of indecision, tacit agreement or having no real semantic function, marked as *mmm, yyy, mhm*; whisper; background noises (e.g. the sounds of various machines or appliances), together with their duration, indicated in square brackets; pauses in speech, with the indication of their duration; overlapping speech and moments of unclear speech (indistinguishable or incomplete words). Besides, the English lexical and semantic borrowings were marked (by means of the use of bold type for lexical borrowings and underlining for semantic borrowings).

6 The results of the study – a brief summary

The results of the study, i.e. the use of English borrowings in informal spoken Polish, will not be discussed in detail, since the aim of present article was to describe the corpus collected for the study as well as to present certain problems associated with collecting a corpus of spontaneous spoken language.

Altogether, 78 English lexical borrowings (including derivatives) were found in the corpus (225 tokens). This number may seem large at first glance, but when it is contrasted with the number of running words of the entire corpus, it becomes evident that English loanwords constitute a very small percentage of the corpus, namely 0.3715%. As for semantic borrowings, 44 types of them (including derivatives) were found in the corpus (158 tokens), which constitutes 0.26088% of the entire corpus [18].

7 Conclusions

Altogether, it turned out, as was expected, that the use of a corpus is not the most efficient way of studying English borrowings in informal speech, since the loans (both lexical and semantic) are not very numerous. What is more, the majority of them were

used only once in the entire corpus. On the other hand, many of them have not been discussed in the literature on the subject so far. Most importantly, however, it would seem that the study can be described as a reliable one: the conclusion is that English borrowings are used relatively infrequently in spoken informal Polish.

In conclusion, it can also be stated that there is still a need to study English borrowings in spoken Polish, as this area is generally neglected. However, the project of this type should not be carried out by an individual linguist, but by a number of them, preferably from different universities (i.e. located in different parts of Poland) so that the informants would be varied as to their place of living, sex, age, occupation, education, social and economic background and knowledge of English and other foreign languages. The base of such a research should be a large corpus of spoken spontaneous language (i.e. forms such as discussions, interviews emitted e.g. by the television should not be included), as it remains the only objective method of studying loanwords in a spoken variety of a language.

References

[1] Crystal, D. (1997). *A Dictionary of Linguistics and Phonetics* (4th edition). Oxford: Blackwell Publishers Ltd.

[2] Dunaj, B. (ed.) (1979). *Wybór tekstów języka mówionego mieszkańców Krakowa*. Kraków.

[3] Dunaj, B. (2000). Nowe słownictwo w leksykografii. In: Mazur, J. (ed.) *Słownictwo współczesnej polszczyzny w okresie przemian*, pages 33–38, Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.

[4] Kamińska, M. (ed.) (1989). *Wybór tekstów języka mówionego mieszkańców Łodzi i regionu łódzkiego. Generacja najstarsza*. Łódź.

[5] Kamińska, M. (ed.) (1992). *Wybór tekstów języka mówionego mieszkańców Łodzi i regionu łódzkiego. Generacja starsza, średnia i najmłodsza*. Łódź.

[6] Lubaś, W. (ed.) (1976). *Miejska polszczyzna mówiona. Metodologia badań.* Katowice.

[7] Lubaś, W. (ed.) (1978). *Teksty języka mówionego mieszkańców miast Górnego Śląska i Zagłębia*, vol. 1. Katowice.

[8] Lubaś, W. (ed.) (1980). *Teksty języka mówionego mieszkańców miast Górnego Śląska i Zagłębia*, vol. 2. Katowice.

[9] Lubaś, W. (1979). *Społeczne uwarunkowania współczesnej polszczyzny. Szkice socjolingwistyczne.* Kraków: Wydawnictwo Literackie.

[10] McEnery, T. and Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

[11] Myrczek, E. (2000). Corpus – its definition, compilation, taxonomy and future. Linguistica Silesiana 21, 43–62.

[12] Otwinowska-Kasztelanic, A. (2000). *A study of the lexico-semantic and grammatical influence of English on the Polish of the younger generation of Poles (19-35 years of age)*. Warszawa: Wydawnictwo Akademickie Dialog.
[13] Otwinowska-Kasztelanic, A. (2000). *Korpus języka mówionego młodego pokolenia Polaków*. Warszawa: Wydawnictwo Akademickie Dialog.
[14] Pisarkowa, K. (1975). *Składnia rozmowy telefonicznej*. Wrocław: Zakład Narodowy im. Ossolińskich.
[15] Poplack, S. and Sankoff, D. (1984). Borrowing: the synchrony of integration. Linguistics 22, 99-135.
[16] Poplack, S., Sankoff, D. and Miller, Ch. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics 26, 47-104.
[17] Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
[18] Zabawa, M. (2006). *English lexical and semantic loans in informal spoken Polish. Vol. I-II. (Vol I: The dissertation; vol. II: The corpus of informal spoken Polish).* Unpublished doctoral dissertation. University of Silesia.
[19] Żydek-Bednarczuk, U. (1994). *Struktura tekstu rozmowy potocznej*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Vybudování databází na základě slovníku jako korpus

Miloud Taïfi[1] a Patrice Pognan[2]

[1] Universita Sidi Mohamed ben Abdallah, Fez, Maroko
[2] INALCO, LALIC-Certal, Paříž, Francie

Abstrakt. Slovník berberštiny středního Maroka (tamazight) obsahuje skoro 7200 kořenů. Automatické vybudování databáze se ukazálo možným díky tomu, že je tento slovník dobře strukturován. Dva moduly jsou spojené s automatickým vyhledáváním a rozšiřováním struktur, třetí s vybudováním databází čtením výsledků. Hlavní aplikace bude příprava francouzsko-berberského slovníku. Francouzský překlad tohoto článku dáme na server `http://www.lalic.paris4.sorbonne.fr`.

První slovník berberštiny středního Maroka (tamazight) byl publikován profesorem Taïfim v roce 1991 jako rukopis. Druhá verze vyjde koncem roku 2009. Byla vyhotovena v rámci marocko-francouzské vědecké kooperace. Jedná se o opravenou a o 50% rozšířenou verzi. Jak to bývá v semitských jazycích, tento slovník je abecedně seřazen podle kořenů (je jich skoro 7200). Slovník se představuje jako souhrn 29 souborů ve Wordu (optické snímání rukopisu selhalo a bývalý slovník se musel ručně naklepat). Je dobře strukturován, jak je z ukázky vidět:

KOŘEN – ♦ slovo – gramatické údaje – ► významy – ● příklady

♦ *abadr (u)*

ibadern ► action de se presser, de se hâter, de se dépêcher ; hâte.

BDR

♦ **badr**

badr, tbadar / tbidir, ur-badr ► démolir, détruire (objet, édifice ...). (v. aussi : *hdem*, hdm ; *rdem*, rdm). ● *ibadr agadir s ugenzim*, il a démoli le mur avec une pioche. ► battre qn. en le jetant par terre, faire tomber qn. en le frappant, en le poussant. ● *iwwet-t s rrkel ibadr-t*, il l'a

♦ *abudr (u)*

ibudren ► maquillage (résultat, état). ● *ur as iġudi ubudr*, elle ne s'est pas bien maquillée.

♦ *abudir (u)*

ibudurn ► fard, produit qui sert à se maquiller (crème, rouge à lèvres ...)

♦ *asbudr (u)*

isbudren ► action de maquiller, de farder ; maquillage (action).

BDR

♦ **abadder** (u)

První etapa spočívala v tom, že se soubory ve Wordu změnily na soubory v Unikódu UTF-8 a spojily dohromady do poměrně velkého korpusu o 2 700 000 znaků. Na základě tohoto korpusu se vytvořily dva seznamy: korespondence berberských slov s kořeny a seznam francouzských významů s odpovídajícími berberskými kořeny.

Třetí a aktuální úkol, stále na základě slovníku, je automatické vybudování databáze berberštiny středního Maroka. Dva moduly jsou spojené s automatickým vyhledáváním struktur, třetí s vybudováním databází:

1 automatické vyhledávání slovníkových struktur
1.1 rozpoznávání aktuálních struktur slovníku
1.2 rozšiřování struktur doplňováním morfologických údajů
2 vybudování databází čtením výsledků z etapy 1.2

1 Automatické vyhledávání slovníkových struktur

1.1 Etapa rozpoznávání struktur slovníku je sama rozdělena do pěti programů. Vychází z takového souboru:

♦ abadr (u)
ibadern ► action de se presser, de se hâter, de se dépêcher ; hâte.
BDR
♦ badr
badr, tbadar / tbidir, ur-badr ► démolir, détruire (objet, édifice ...). (v. aussi : hdem, hdm ; rdem, rdm). ● ibadr agadir s ugenzim, il a démoli le mur avec une pioche. ► battre qn. en le jetant par terre, faire tomber qn. en le frappant, en le poussant. ● iwwet-t s rrkel ibadr-t, il l'a frappé avec un coup de pied et l'a fait tomber.
♦ Tu — tubadr
tubadr, ttubadar, ur-tubadr ► être démoli, détruit (objet, édifice ...). ► être mis à terre (personne).
♦ abadr (u)
ibadren ► action de démolir, de détruire ; démolition, destruction. ► action de mettre qn. à terre en le frappant, en le poussant.
BDR
♦ budr
budr, tbudur, ur-budr ► se maquiller, se farder ; se faire beau ; soigner son apparence. (v. aussi :

1. První program čte základní soubor v Unikódu, určuje kořeny podle písma (velká písmena s diakritickými znaménky) a rozděluje jednotlivá hesla prázdným řádkem. Dále začíná rozeznávat strukturu a označuje ji určitým počtem tabelací (jedna tabelace pro každou úroveň). Tady určuje úroveň slov.
2. Druhý program jde hlouběji do struktury a určuje významy.
3. Třetí určuje příklady.
4. Tento program provádí změny, které budou během druhé etapy nutné pro automatické generování morfologických údajů: dává do jednoho záznamu slovo a základní morfologické údaje.
5. Poslední program lehce upravuje úroveň kořene: přidává k němu odkazy na případnou přítomnost kořene v jiných jazycích.

Výsledky první etapy (1.1) jsou uvedeny níže:

BDR
- ◆ badr, badr, tbadar / tbidir, ur-badr
  - ► démolir, détruire (objet, édifice ...). (v. aussi : hdem, hdm ; rdem, rdm).
    - ● ibadr agadir s ugenzim, il a démoli le mur avec une pioche.
  - ► battre qn. en le jetant par terre, faire tomber qn. en le frappant, en le poussant.
    - ● iwwet-t s rrkel ibadr-t, il l'a frappé avec un coup de pied et l'a fait tomber.
- ◆ Tu — tubadr, tubadr, ttubadar, ur-tubadr
  - ► être démoli, détruit (objet, édifice ...).
  - ► être mis à terre (personne).
- ◆ abadr (u), ibadren
  - ► action de démolir, de détruire ; démolition, destruction.
  - ► action de mettre qn. à terre en le frappant, en le poussant.

BḤŠ ar., kb.
- ◆ abuḥšiš (u), ibuḥšišn
  - ► gorge, gosier (œsophage et trachée artère).
    - ● iqqur-as ubuḥšiš, il a la gorge sèche.
  - — bu-ubuḥšiš, fém. : mm-ubuḥšiš
  - ► personne criarde, qui parle fort.
- ◆ tabuḥšišt, tibuḥšišin
  - ► dim. du précéd.

1.2 Rozšiřování struktur doplňováním morfologických údajů

Druhá etapa má za úkol automaticky vypočítat další morfologické hodnoty na základě původní struktury. Je rozdělena do tří programů seřazených podle jevů a jejich složitosti.

1. V berberštině existuje velká homografie kořenů. Pro přehlednost a pro zajištění správného fungování databází bylo nutno očíslovat kořeny. Tak např. kořen „BDR“ uvedený výše bude označen jako „BDR BDR3“.

2. Vyhledávání lexikálních kategorií a obohacování morfologických údajů (gramatémů): proveditelnost projektu čistě závisí na tomto programu. Kdyby býval neprovedl správnou analýzu lexikálních kategorií, byla by konstrukce databází nemožná.

Neurčuje jenom „třídy“ (lexikální kategorie), ale určuje taky „kmeny“ (typ odvození: verbo-nominální nebo nominální derivace). Pojem rozdělení lexika na dvě části, verbo-nominální a nominální, je v lingvistice hamito-semitských jazyků běžný. Aplikujeme ho i na indoevropské jazyky.

Program rozeznává:

- jednoduchá slovesa označená jako verbo-nominální derivace (VN), sloveso (V) a jednoduché sloveso (1.1). Číselná škála (1.1. - 1.2. - 1.3. - 2., přitom „1“ odkazuje na verbo-nominální derivaci a „2“ na nominální derivaci) umožňuje automatické seřazení slov pod kořeny u výstupů z druhého typu databází (viz níže 2.2.1).

- odvozená slovesa: „VN V 1.2“

- deverbativa: „VN SUBST 1.3.“

- a konečně „nomen“ (podstatná a přídavná jména) s nominální derivací (N) s hodnotami (SUBST nebo ADJ) a (2.): „N SUBST 2“.

U všech sloves (jednoduchých a odvozených) jsou ještě uvedeny čtyři hlavní časy, které pak umožní automatické časování uvnitř databáze. To jsou (aorist, dokonavý tvar [také často nazývaný „préteritum“], nedokonavý tvar, negativní dokonavý tvar) např.:

sloveso „dělat“ : „sker sker teskar ur-skir“.

U všech nominálních forem (deverbativa a substantiva) se generuje rod podle formy slova: až na výjimky, často odvozené z arabštiny, je berberský „nomen“ pravidelný. V kategorii substantiv slova začinající na „a/i/u“ jsou mužského rodu, slova začinající na „t“ jsou ženského rodu (a většinou končí též na „t“). Střední rod neexistuje. Dále se generují tzv. „vázané“ formy v jednotném a v množném čísle na základě „volných“ forem. U přídavných jmen se tyto formy generují pro mužský a ženský rod.

3. Příklady představují nejhlubší úroveň struktury slovníku. Vedle příkladu obsahuje ještě jejich záznam překlad a eventuelně doslovný překlad v závorkách. Struktura záznamu vypádá takto:

• irḍel a-s leflus (litt. il a prêté à lui de l’argent), il lui a prêté de l’argent.

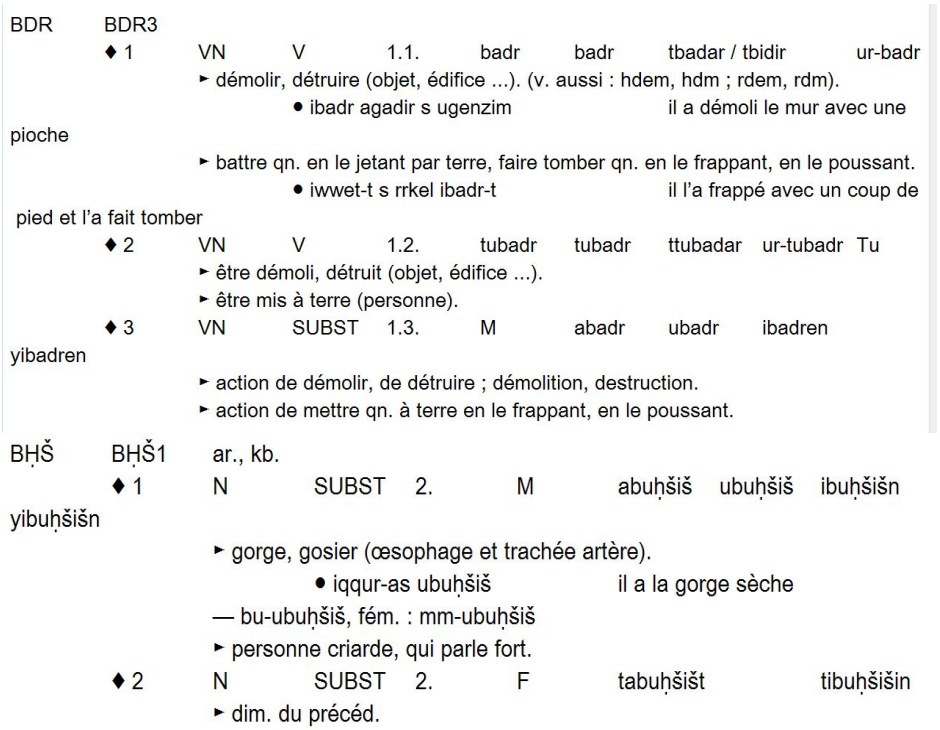
„• berberský příklad (překlad doslova) , překlad .“

Typický oddělovač je čárka, která odděluje příklad a doslovný překlad v závorkách na jedné straně a korektní překlad na straně druhé. Problém je v tom, že čárka je špatným oddělovačem proto, že se může navíc objevovat ve všech třech částech záznamu, a to několikrát. Z toho vyplývá, že odpovídající algoritmus je poměrně složitý. Nedá se vyloučit, že některé záznamy nebudou plně nebo vůbec vyřešeny. Jeden z použitých postupů je rozdělení na segmenty ohraničené čárkou a otestování přítomnosti v segmentu buď typických berberských znaků nebo typických francouzských znaků, a to tak, že berberské znaky se otestují zepředu od druhého segmentu a francouzské naopak zezadu od předposledního segmentu. Bohužel se může stát, že jeden segment (ba i několik) uprostřed záznamu nemá vůbec žadný zvláštní znak!

Zvláštní berberské znaky jsou: „ č š ž ḍ ġ ḥ ṛ ṣ ṭ ẓ ɛ “ a francouzské jsou: „ ’ à æ é è ê ï î ô œ ù û ç “.

Uvažujeme také o tom, že budeme vyhledávat v segmentech případnou přítomnost berberských předložek, spojek a částic jako např. „d“, „n“, „s“ (má stejný význam jako v češtině a slovenštině!), …

Výsledky druhé etapy (1.2) vypadají takto:



2 Vybudování databází čtením výsledků z etapy 1.2

Připravili jsme a otestovali dva prototypy. První byl připraven tak, aby odpovídal struktuře:

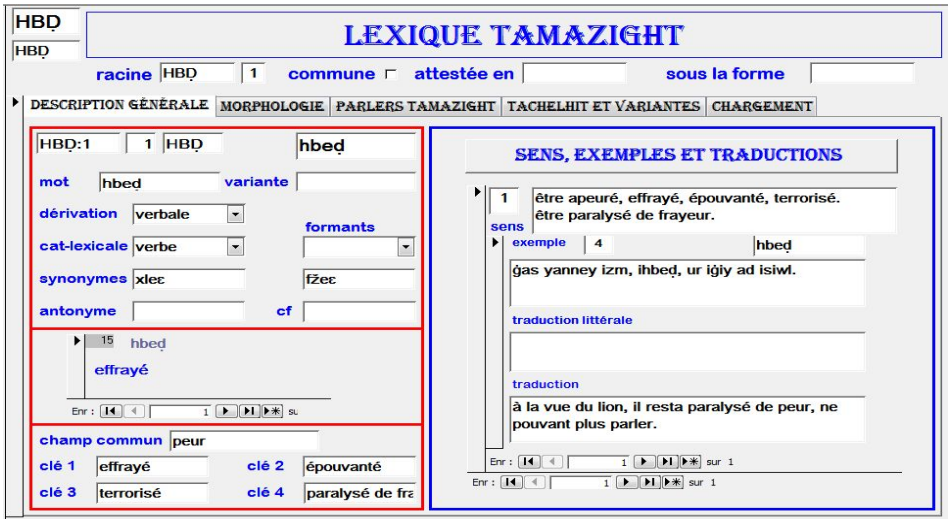
KOŘEN – ♦ slovo – gramatické údaje – ► významy – ● příklady.

Druhý má takovou strukturu, kde nejvyšší úroveň je úrovní slova se zmínkou kořene:

slovo a kořen – lexikální kategorie a gramatické údaje – významy – příklady. Načteme připravené údaje z etapy 1.2 do těchto dvou prototypů, ale pravděpodobně budeme nadále pracovat s druhým prototypem, protože odpovídá naší nynější koncepci databází.

2.1 První prototyp

Jde v podstatě o přesný obraz slovníku:

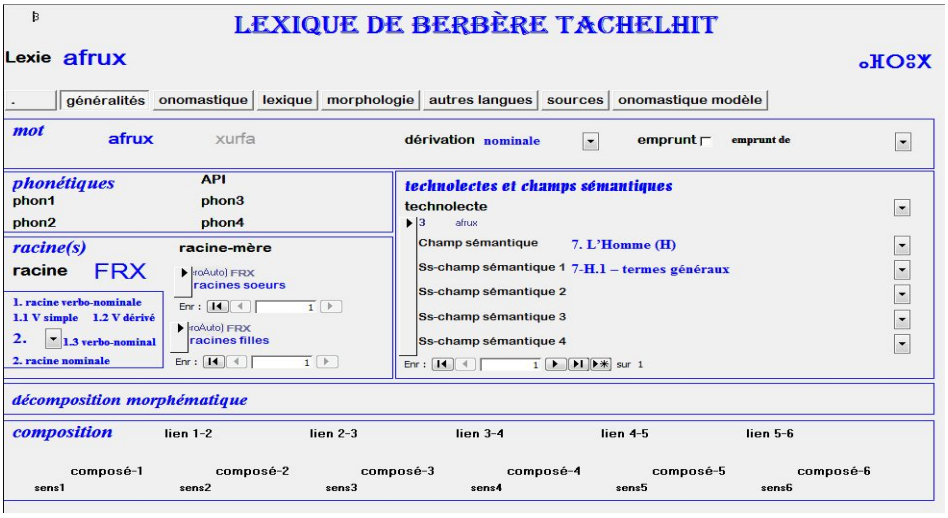


2.2 Druhý prototyp

Po vybudování několika databází jsme nabyli přesvědčení, že konstrukci databází do nekonečna opakovat nelze. Proto jsme se rozhodli definovat prototyp, který by odpovídal našim požadavkům. Vybudovali jsme prototyp databáze, který je jedinečný, víceúčelový a vícejazyčný.

Tento prototyp odpovídá nejširšímu popisu lingvistického systému indo-evropských a hamito-semitských jazyků. Jednoduchou kopií jsme z něho už dostali několik databází: jednak pro češtinu, slovenštinu, slovinštinu a albánštinu, jednak pro berberštinu. Rýsuje se i projekt rozpracování všech druhů maghrebské arabštiny. Prototyp nyní rozšiřujeme na struktury tureckých jazyků hlavně na základě turečtiny a uzbečtiny. Je to důležité proto, že se tím otevírají další možnosti směrem ke všem aglutinačním jazykům, jako jsou třeba ugrofinské jazyky, baskičtina a korejština...

Víceúčelovost prototypu je dána tím, že umožňuje shromáždit lexikální znalosti, které vyústí ve výrobu slovníků, gramatické znalosti ze všech úrovní lingvistického popisu, ať jsou synchronické nebo diachronické, a odborné znalosti (v technických, vědeckých a lékařských oborech). Tato víceúčelovost se odráží v přítomnosti několika druhů struktur.



2.2.1 Lexikální struktura je hlavní a centrální strukturou databázového systému. Je určena především pro výrobu všelijakých rejstříků a slovníků. Podrobný popis se najde v [15] a [16]. Obsahuje tyto úrovně:

1. slovo a kořen (tento typ databáze má třídění podle slov). Hodnoty (1.1, ... až 2.) vygenerované během etapy 1.2 mají za úkol správné třídění slov u každého kořene, pokud chceme dostat slovník tříděný podle kořenů.
2. lexikální kategorie (může jich být víc)
3. významy
4. příklady a další...

2.2.2 Přilehlá struktura morfologických údajů

Obsahuje informace o cizím původu, o neologii, o morfématickém rozboru s udáním kořene a dalších afixů. Podává též všeobecné údaje o slovesech, podstatných a přídavných jménech.

2.2.3 Vedlejší struktury

Jsou trojího druhu. Jedná se pokaždé o nový směr aplikací.

- tabulky pro časování sloves a skloňování nominálních forem. Mají velkou platnost pro výuku berberštiny (nejenom v cizině pro cizince a pro děti emigrantů, ale i v maghrebských školách a sdruženích).
- tabulky pro srovnávání jazyků stejné skupiny. Typickým příkladem je srovnávání
- západních slovanských jazyků, slovinštiny a ruštiny na základě diachronické fonologie. V berberštině se srovnávací práce týkají hlavně, ba výlučně synchronického stavu jazyka (výjimku tvoří jedině Allati [2]). Zato existuje až příliš velká bohatost dialektů a místních variant.
- tabulky pro popis klasifikačních oborů:

Tato složka se týká vědeckých, technických a lékařských oborů, kde klasifikace hraje velkou roli. V prototypu jsou předem definovány dvě stálé tabulky: latinská

a francouzská. Při budování prototypu jsme tuto složku definovali na základě říše „funghi“, totiž hub, které již nepatří do rostlinné říše a které se staly samostatnou říší. Slova, která patří do těchto oborů, budou mít vedle normálního lingvistického popisu ještě odborný popis v rámci příslušné tabulky. Bude to významná pomoc při tvorbě terminologických slovníků.

2.3 Automatické vytvoření a rozšíření databáze

2.3.1 Automatické vytvoření databáze

Pokud chceme automaticky vytvořit databázi podle prvního prototypu, stačí si programem načítat výsledný soubor (etapa 1.2) podle vzoru potřebných tabulek v předurčeném pořadí:

KOŘEN – ♦ slovo – gramatické údaje – ► významy – • příklady.

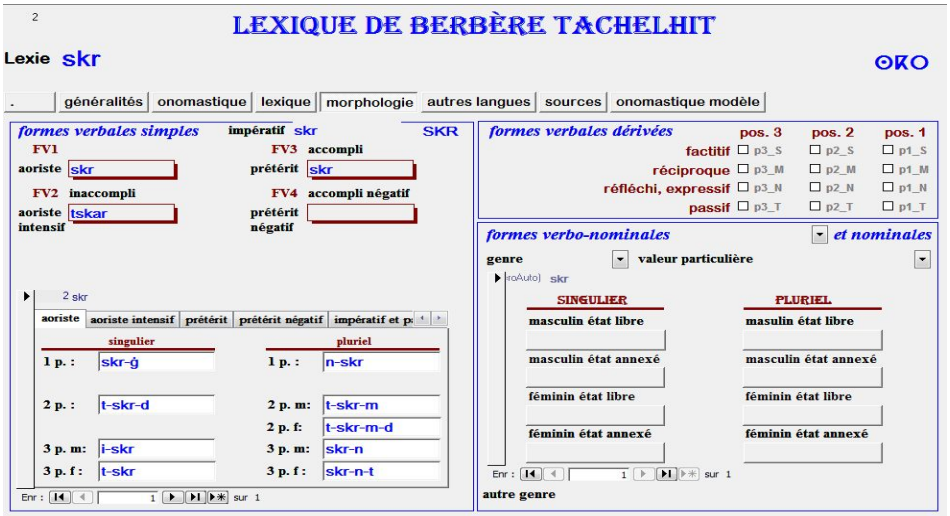
Oproti tomu je nutno pro výstavbu databáze podle druhého prototypu smísit dva první záznamy do jednoho záznamu, a to následujícím způsobem:

```
#--- 1. table du niveau mot - racine
badr    BDR3♦badr1    BDR    BDR3    1.1.    VN    V        badr    badr    tbadar / tbidir    ur-badr
tubadr  BDR3♦tubadr2  BDR    BDR3    1.2.    VN    V  Tu    tubadr  tubadr  ttubadar ur-tubadr
abadr   BDR3♦abadr3   BDR    BDR3    1.3.    VN    SUBST M  abadr   ubadr   ibadren  yibadren
```

Dále je způsob čtení dalších tabulek u obou modelů stejný.

2.3.2 Interní programování databáze

Stejně jako v projektu pro slovanské jazyky je jedním z hlavních bodů automatické časování a skloňování vedené v rámci databáze. Výsledky jsou pak vloženy do příslušných tabulek.



Na příkladu posledního formuláře si ukažeme postup naplňování databáze programováním. Např. pro záznamy „badr“ a „tubadr“ dává program čtení základní hodnoty, respektive „badr – badr – tbadar / tbidir – ur-badr“ a „tubadr – tubadr – ttubadar – ur-tubadr“, do připravených polí s názvem FV1 [forme verbale 1] až FV4 [forme verbale 4] (klasičtější pojmenování jsou vedle polí nebo nad nimi). Teprve na základě těchto informací se automaticky časuje celé sloveso (to se dělá v programu Access pomocí objektově orientovaného programovacího jazyka VBA) a výsledky jsou automaticky vloženy do příslušné vedlejší tabulky (viz předešlý formulář).

2.3.3 Obsahové rozšíření

Databáze, která se těmito procesy vytvoří, bude dále rozšířena dalšími vědeckými pracemi, hlavně universitními doktoráty různých autorů, ke kterým patří mimo jiné Ameur, Amrani, El Mandour, Jarmouni,...

Využití databáze

Tato databáze bude mít různé aplikace v oboru lexikálních a gramatických studií. Nejdůležitější aplikací bude příprava tvorby francouzsko-berberského slovníku. Vyvinuli jsme proceduru, která dovede obrátit celý obsah databáze tak, že autor slovníku z francouzštiny do cizího jazyka má k dispozici veškerý materiál předešlého slovníku (z cizího jazyka do francouzštiny). Tento postup nemá žádný vliv na koncepci slovníku, ale zato tvorbu slovníku značně urychluje. Touto procedurou jsme poměrně rychle vytvořili francouzsko-slovenský slovníček v rámci evropského projektu na vypracování vyučovací metody slovenštiny [5].

Najdou se i další aplikace jako je gramatika berberštiny a metoda výuky berberštiny ve francouzštině. Věříme, že taková databáze bude pro výuku berberštiny velmi užitečná.

Pokud nám to vyjde, dáme konečný výsledek na server pro veřejné použití.

Literatura

[1] Achab, R. (1996). *La néologie lexicale berbère (1945-1995)*. Peeters, Louvain.
[2] Allati, A. (2002). *Diachronie tamazighte*. Université de Tétouan.
[3] Ameur, M., Bouhjar, A., Boukhris, F., Boukous, A., Boumalk, A., Elmedlaoui, M., Iazzi, E. M., Souifi, H. (2004). *Initiation à la langue amazighe*. Publications de l’IRCAM, Rabat.
[4] Ameur, M., Boumalk, A. (2004). *Standardisation de l’amazighe*. Publications de l’IRCAM, Rabat.
[5] Baranová, E., Křečková, V., Lemay, D., Pognan, P. (2007). *Découvrir et pratiquer le slovaque*. Projet européen ALPCU [lingua II], L’Asiathèque, Paris.
[6] Bentolila, F. (1981). *Grammaire fonctionnelle d’un parler berbère. Aït Seghrouchen d’Oum Jeniba*, Selaf, Paris.

[7] Boumalk, A. (2003). *Manuel de conjugaison du tachelhit*, L'Harmattan, Paris.
[8] Chaker, S. (1995). *Linguistique berbère. Etudes de syntaxe et de diachronie*, Peeters, Louvain.
[9] Dallet, J.-M. (1982). *Dictionnaire Kabyle – Français*, Selaf, Paris.
[10] El Mountassir, A. (2003). *Dictionnaire des verbes tachelhit – français. (parler berbère du sud du Maroc)*, L'Harmattan, Paris.
[11] Mammeri, M. (1992). *Précis de grammaire berbère*. Editions EnAP et Inna yas, Alger et Tizi-Ouzou.
[12] Naït-Zerrad, K. (1995). *tajeṛṛumt n tmaziγt tamirant (taqbaylit)*. ENAG, Alger.
[13] Naït-Zerrad, K. (1998 -). *Dictionnaire des racines berbères (formes attestées)*. Peeters, Louvain.
[14] Naït-Zerrad, K (2004). *Linguistique berbère et applications*, L'Harmattan, Paris.
[15] Pognan P. (2009). Définition d'une base de données linguistiques "à tout faire" comprenant le français – comme langue pivot – et les langues slaves de l'Ouest. In *Congrès international Studia Romanistica Beliana*, Banská Bystrica.
[16] Pognan P. (2009). Définition d'un prototype général de bases de données (étude des langues slaves de l'Ouest dans une visée multilingue). In *Meta-language and Encoding Scheme Design for Digital Lexicography – Innovative Solutions for Lexical Entry Design in Slavic Lexicography, MONDILEX Third Open Workshop*, Académie des Sciences Slovaque, Bratislava.
[17] Sadiqi, F. (2004). "*A Grammar of Amazigh*", Université Sidi Mohamed ben Abdellah-PARS, Fez.
[18] Taïfi, M. (1991). *Dictionnaire tamazight – français. Parlers du Maroc central*, L'Harmattan-Awal, Paris.
[19] Taïfi, M. (2006). Si les berbérophones ne s'entendent pas oralement, qu'ils s'écrivent ! Pour une écriture grammaticale du berbère. In *Standardisation de l'amazighe*. Publications de l'IRCAM, Rabat.
[20] Taïfi, M., Pognan, P. (2005). Langues berbères: à la recherche du système perdu. In *Colloque international « Linguistique amazighe: les nouveaux horizons »*. Tétouan.
[21] Taïfi, M. (připravuje se). Dictionnaire raisonné berbère – français. Parlers du Maroc.

Appendix

creative commons

Attribution-ShareAlike 3.0 Unported

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

a. **"Adaptation"** means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

b. **"Collection"** means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.

c. **"Creative Commons Compatible License"** means a license that is listed at http://creativecommons.org/compatiblelicenses that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.

d. **"Distribute"** means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

e. **"License Elements"** means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

f. **"Licensor"** means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

g. **"Original Author"** means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

h. **"Work"** means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

i. **"You"** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

j. **"Publicly Perform"** means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

k. **"Reproduce"** means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

b. to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";

c. to Distribute and Publicly Perform the Work including as incorporated in Collections; and,

d. to Distribute and Publicly Perform Adaptations.

e. For the avoidance of doubt:

i. **Non-waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme

cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

ii. **Waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

iii. **Voluntary License Schemes**. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.

b. You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US)); (iv) a Creative Commons Compatible License. If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the "Applicable License"), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of each Adaptation You Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

c. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with

the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Section 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

d. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

b. Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

c. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

d. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

e. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

f. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

NLP, Corpus Linguistics, Corpus Based Grammar Research

Editors Jana Levická Radovan Garabík

Cover Design by Vladimír Benko
Typeset by Marek Ivančík

Printed by Tribun EU s.r.o.
Gorkého 41, 602 00 Brno, Czech Republic
http://www.librix.eu

ISBN 978-80-7399-875-2

First published by Tribun EU
Brno 2009