Д. В. Сичинава
Институт русского языка им. В. В. Виноградова РАН;
НИУ «Высшая школа экономики»
(Россия, Москва)
mitrius@gmail.com

ПАРАЛЛЕЛЬНЫЕ ТЕКСТЫ В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: НОВЫЕ ЯЗЫКИ И НОВЫЕ ЗАДАЧИ*

В статье рассказывается об основных направлениях пополнения и содержательного развития параллельных корпусов НКРЯ за 2015–2019 гг. В разделе «Новые языки» речь идёт о новых языковых парах, возникших за этот период, об архитектуре и разметке соответствующих корпусов. По сравнению со списком языков, образующих двуязычные параллельные пары с русским и доступных в 2015 году, в НКРЯ появились следующие новые языки: башкирский, бурятский, китайский, литовский, финский, чешский, шведский. Продолжается традиция создания параллельной части НКРЯ при помощи ряда автономных российских и зарубежных команд, координирующих свои усилия с группой разработчиков Корпуса в Москве. Практически все новые языки ставили перед разработчиками Корпуса те или иные особые задачи, связанные с их морфологической или иной пословной разметкой. Существенно вырос за четыре года объём также некоторых из уже доступных в 2015 г. языковых пар.

В разделе «Новые задачи» раскрываются основные содержательные направления, разрабатываемые в рамках разных языковых корпусов — региональное разнообразие языка, жанровое разнообразие стилей, расширение функциональности и типов разметки и др. На настоящем этапе, как благодаря увеличению типологического разнообразия задействованных языков и письменностей, так и благодаря использованию более сложных морфологических анализаторов, существенно расширился набор дополнительных параметров разметки, поиск по которым может быть релевантен для работающего с параллельным корпусом. Цель включения в корпус образцов полицентричных языков стала одной из важнейших. Жанровое

* Статья написана (и большинство обсуждаемых корпусов разработано) при поддержке проекта РФФИ № 17-29-09154 «Динамика языковой системы: корпусное исследование синхронной вариативности и диахронических изменений в текстах разных типов» (руководитель Г. И. Кустова).

разнообразие корпусов, в 2015 году лишь намечавшееся, в 2019 году является одной из главных целей, причем эта цель учитывается с самого начала создания новых языковых пар.

В статью включен отдельный исследовательский сюжет — изучение плюсквамперфекта по многоязычному корпусу. Анализ многоязычного текста по материалам расширенной коллекции позволяет построить сеть расстояний между данными для граммемы плюсквамперфекта в 24 идиомах Европы.

Ключевые слова: параллельные корпуса; двуязычные корпуса; многоязычные корпуса; разметка; репрезентативность; плюсквамперфект

Со времени публикации предыдущей статьи о параллельных (включая многоязычный) (под)корпусах в составе Национального корпуса русского языка [Сичинава 2015] архитектура и функциональность параллельных корпусов в составе НКРЯ не претерпели радикальных изменений. При этом существенно увеличился объём этих корпусов: на май 2019 г. в поиске доступно 98,2 млн словоупотреблений — близко к психологически важной отметке в 100 млн — по сравнению с 70 млн в 2015 г.; таким образом, соответствующий раздел НКРЯ становится одним из крупнейших лингвистических параллельных корпусов, наряду с InterCorp в составе Чешского национального корпуса (http://ucnk.korpus.cz/intercorp/). Появились новые языки, представленные в двуязычных парах, появились новые задачи, связанные как с разметкой корпусов, так и с функциональностью поиска.

Существенно вырос за четыре года также объём некоторых из уже доступных в 2015 г. языковых пар:

Язык	**Текущий объём, обе языковые пары, словоформы на обоих языках, млн словоупотреблений**	**Прирост объёма по сравнению с 2015 г.**
Английский	28,3	15%
Белорусский	10,9	63%
Испанский	2,4	88%
Итальянский	4,8	21%
Латышский	3,0	324%
Немецкий	9,7	27%
Французский	5,1	120%
Эстонский	0,6	51%

В дальнейшем, как и в работе [Сичинава 2015], сокращенное обозначение «L-й корпус», где L-й — название языка, используется как практическая замена для более громоздкого сочетания «параллельные L-русский и русско-L-й корпус».

1. Новые языки

По сравнению со списком языков, образующих двуязычные параллельные пары с русским и перечисленных в [Сичинава 2015], в НКРЯ появились следующие новые языки: башкирский, бурятский, китайский, литовский, финский, чешский, шведский. Совокупный объём только этих новых корпусов составляет

12 млн размеченных словоупотреблений, или токенов, причем более половины этого количества приходится на шведский корпус, благодаря его главной разработчице Н. В. Перковой вошедший в пятерку самых крупных параллельных корпусов НКРЯ по объёму. В разработке этих корпусов активно принимали участие специалисты, работающие на территории распространения соответствующих языков. Таким образом, продолжается традиция создания параллельной части НКРЯ при помощи ряда автономных российских и зарубежных команд, координирующих свои усилия с группой разработчиков Корпуса в Москве. Во многих случаях речь идёт о тесном сотрудничестве с командами уже существующих одноязычных корпусов, национальных или иных, об обмене текстами, а также и об интеграции с другими «семействами» параллельных корпусов, например, для чешского или финского. Практически все новые языки ставили перед разработчиками Корпуса те или иные особые задачи, связанные с их морфологической или иной пословной разметкой. Отметим, что почти во всех указанных случаях (кроме чешского) речь идёт о «внешней» разметке по отношению к базовой используемой в Корпусе морфологической разметке «Яндекса», а затем о той или иной ее адаптации под используемый формат Корпуса. Здесь также естественным было заимствование тех или иных решений из существующей практики одноязычных корпусов вместе с унификацией грамматических обозначений, которую диктует общая архитектура параллельных корпусов НКРЯ.

1.1. Башкирский корпус

В состав НКРЯ включены параллельные башкирско-русские и русско-башкирские тексты, подготовленные в 2016 г. под общим руководством Б. В. Орехова командой разработчиков-волонтёров из Башкирии для компании «Яндекс», совокупным объёмом 550 тысяч слов. Для включения в Корпус эта коллекция потребовала некоторой доработки, в частности, создания хотя бы упрощенной метатекстовой разметки (в первоначальной версии она отсутствовала, а также было широко представлено как слияние, так и разделение изначальных текстов). Все файлы снабжены башкирской морфологической разметкой, разработанной Б. В. Ореховым [2014], а также переведены им же специально для НКРЯ в принятый в корпусе формат XML.

Основная масса текстов — новости и статьи из двуязычных газет и электронных изданий 2010-х годов, но широко представлены и другие жанры. В частности, впервые в практике НКРЯ в корпус включаются нетекстовые (при этом изначально именно двуязычные параллельные) жанры — башкирско-русский словарь и разговорник. Это способствует резкому повышению охвата представленных в корпусе лексики, фразеологии и конструкций, хотя, безусловно, подобные источники имеют скорее пограничный статус между текстом как таковым и базой данных. Ср. полезное для диахронического исследования лексики и фразеологии включение в украинский одноязычный корпус ГРАК [Шведова та ін. 2017–2019] полностью «Словаря украинского языка» Б. Д. Гринченко и «Русско-украинского словаря

устойчивых выражений» И. О. Виргана и М. М. Пилинской, которые, между прочим, тоже являются двуязычными — толкования или соответствия украинской лексики и фразеологии даются в них по-русски. Отметим также, что функция «словарь как корпус» позволяет искать внутри словосочетаний и текстовых примеров из статей по самым разным параметрам, а не только по вокабуле статьи или даже отдельным словам из ее текста.

Кроме того, в башкирский корпус входят, полностью или в отрывках, художественные произведения (в том числе изначально написанные на третьем языке, но переведенные на башкирский с русского, например, «Маленький принц» Сент-Экзюпери), официально-деловые тексты (Конституция Башкирии или тексты с сайта Курултая) и развернутые научно-популярные тексты из Википедии, весьма точно переложенные с русского: *Гассман Сальериҙы танылған опера либреттоһы оҫтаһы, һарай шағиры Пьетро Метастазио менән таныштыра, уның йортонда Вена интеллектуалдары һәм артистары йыйылыр була…* (из имеющей статус «избранной статьи» в обеих языковых версиях биографии Антонио Сальери). Кроме Википедии и художественных произведений, направление перевода в этой коллекции текстов не всегда очевидно, что, разумеется, является обычной ситуацией для двуязычных сайтов вообще и СМИ в частности, особенно на постсоветском пространстве. Таким образом, разбиение этого корпуса на башкирско-русскую и русско-башкирскую параллельные части носит в значительной степени условный характер.

1.2. Бурятский корпус

Создание бурятского корпуса поддерживалось специальным проектом РФФИ №15-46-04417 («Бурятско-русский параллельный корпусный модуль») и осуществлялось совместно с Институтом монголоведения, буддологии и тибетологии РАН (Улан-Удэ) под руководством Л. Д. Бадмаевой. В этот корпус объёмом в 400 тыс. словоупотреблений вошли только художественные тексты: один переведенный с русского (пушкинская «Капитанская дочка») и несколько произведений довоенной и послевоенной бурятской литературы XX в., включая один из ключевых текстов бурятской культуры «Путь праведный» Б. Дандарона; авторство, текстология и история перевода этого романа представляют собой отдельную исследовательскую проблему.

Важным «вызовом» для создателей этого корпуса являются достаточно вольные переводы, в том числе с пропусками, что, опять-таки, характерно для традиции художественного перевода с «языков народов СССР» и на них; в значительном проценте случаев находимые примеры не имеют точных соответствий на другом языке. Тем не менее подобным образом выровненные тексты имеют большую ценность как для истории перевода, так и для лингвистического анализа. Тексты получают морфологическую разметку, разработанную Т. А. Архангельским и принятую в одноязычном бурятском корпусе (http://web-corpora.net/BuryatCorpus); большинство лемм снабжено также русским переводом с опорой на специально оцифрованный бурятско-русский словарь.

1.3. Китайский корпус

Создание, поддержка и развитие китайского корпуса, насчитывающего в настоящее время 280 тыс. токенов (число уникальных слов меньше, поскольку в эту сумму входят как разделенные на «слова»-сегменты китайские тексты в оригинальной графике, так и дублирующая их строка транслитерации), безусловно, является одной из самых сложных задач, стоящих перед коллективом НКРЯ. Это диктуется, прежде всего, спецификой китайского иероглифического письма, требующего особой автоматизации процессов словоделения, транскрипции и выравнивания.

В его разработке участвует коллектив сотрудников и студентов из российских (ВШЭ, РГГУ, МГУ, РАНХиГС, Алтайский государственный университет) и китайских вузов; в частности, ключевые роли в проекте играют Л. С. Холкина, К. И. Семенов, О. Р. Валиулин, С. П. Дурнева, М. Н. Якубов. Поиском и предоставлением данных, наряду с выравниванием текстов, занимаются Синь На (и коллектив под руководством Е Цисуна) Института лексикографии Хэйлунцзянского университета, а также Юань Мяосюй (и коллектив под руководством Ван Юн) Института иностранных языков Чжэцзянского университета.

В настоящее время в поиске НКРЯ представлены лишь художественные произведения (три рассказа Лу Синя, русская и советская классика — романы при этом представлены в отрывках), однако выровнены и готовятся к вывеске в Корпусе также тексты разнообразных жанров: это корпус деловой переписки, разработанный К. А. Ульяновой, Евангелие в различных, относящихся к разным историческим периодам и конфессиям, переводах на китайский, выровненное с русским синодальным переводом, учитывавшимся при создании части этих переводов, а также другие художественные тексты: Пушкин, Гоголь, Тургенев, Чехов, Горький, современные китайские авторы — Мо Янь, Лю Чжэньюнь, Юй Хуа.

В настоящее время для выравнивания текстов и словоделения используется коммерческая утилита Skuuper (с некоторым усовершенствованием ее механизма, разработанным М. Н. Якубовым и другими участниками). В текстах, представленных в настоящее время в поиске НКРЯ, использовался «жадный» алгоритм словоделения, при котором сегментирование осуществлялось слева направо автоматически на основании наиболее длинной цепочки иероглифов, доступной как словарный вход. Это давало в принципе приемлемый результат (по данным М. Н. Якубова [2017] и В. А. Морозовой [2018] — 79%), но иногда приводило к неверным анализам, при которых алгоритм при движении по тексту слева направо «попадал» в более длинное и менее вероятное в тексте редкое слово. Например, в рассказе Чехова «Человек в футляре» трижды выделяется имеющееся в словаре слово 的姐 [dījiě] ‘таксистка’, поскольку частотный иероглиф 的, выступающий в подавляющем большинстве случаев с чтением [de] как показатель атрибутивной связи, имеет также гораздо более редкое чтение [dī] ‘такси’, которое объединяется с первым иероглифом идущего далее в тексте слова 姐姐 [jiějie] ‘старшая сестра’.

Для словарной разметки используется свободно доступный электронный китайско-английский словарь CEdict. В статьях этого словаря выделены разные поля для разметки — транскрипция, толкование и требуемый показатель класса (счетное слово); ряд иероглифов, имеющих грамматические значения, размечен дополнительно как грамматические показатели. На основании словарной транскрипции китайский текст продублирован в виде строки сплошной транскрипции с указанием всех точек, где чтение теоретически может быть неоднозначно. Соответствующий алгоритм разметки разработан Е. А. Кузьменко и усовершенствован М. Н. Якубовым (ср. также [Якубов 2017]).

Принципиальной сложностью такой разметки является неоднозначность словоделения и транскрипций и высокая степень омонимии и полисемии: для большинства слов, состоящих из одного иероглифа, и для значительной части более длинных приводятся все альтернативные значения, представленные в словаре (включая крайне малочастотные). Отдельной проблемой является разметка транскрибируемых иероглифами имен собственных (и тем самым само выделение этих слов в тексте), далеко не все из которых имеются в словаре — хотя набор включенных в CEdict имен собственных, в том числе даже названий знаменитых текстов, велик, фамилии русских литературных персонажей или русские микротопонимы, разумеется, в нем отсутствуют. Ниже эта разметка будет разобрана несколько подробнее с точки зрения представленной в ней информации (раздел 2.1)

1.4. Литовский корпус

Литовский корпус (как и латышский[1], шведский, эстонский и финский) представляет собой часть подпроекта по созданию параллельного корпуса с участием литературных языков циркумбалтийского ареала [Sitchinava, Perkova 2019]; об этом ареале с лингвистической точки зрения см. [Dahl, Koptjevskaja-Tamm (ed.) 2001]. Выбор, подготовка и выравнивание текстов осуществлены Н. В. Перковой; морфологическая аннотация корпуса основывается на системе онлайн-разметки, разработанной в университете Витовта Великого в Каунасе [Rimkutė, Daudaravičius, Utka 2007]. Объём корпуса, доступного в Интернет-поиске — 550 тысяч словоупотреблений, пополнение продолжается. Корпус и его подготовленное пополнение включают литовские художественные тексты XX — начала XXI в. (Вайжгантас, А. Шкема, И. Мерас, К. Сая, Ю. Апутис, Р. Гавялис, М. Ивашкявичюс), а также переводы с русского (Чехов, Хармс, Бунин). Параллельный корпус НКРЯ (вместе с некоторыми нехудожественными текстами, например, эссе А. Венцловы) используется в разработанном О. Н. Ляшевской литовском модуле синтаксического корпуса Universal Dependencies (https://universaldependencies.org/treebanks/lt_hse/index.html).

[1] Латышский корпус (о котором см. [Perkova, Sitchinava 2016]), как уже указывалось выше, — наиболее значительно пополнившаяся за последние 4 года из существующих двуязычных пар (более чем вчетверо).

1.5. Финский корпус

Финский корпус НКРЯ развивается в тесном сотрудничестве с Тамперским университетом и проектом FinCLARIN, в рамках которого уже разработаны параллельный финско-русский и русско-финский корпуса ПарФин и ПарРус [Михайлов, Хярме 2015]. Стороны обмениваются текстами с целью избежать дублирования работы, кроме того, в проекте НКРЯ используется открытый морфологический анализатор OMorfi (https://github.com/flammie/omorfi), используемый и в проектах университета Тампере.

В финский корпус НКРЯ (подбором текстов, сканированием, распознаванием, вычиткой и выравниванием занимается К. О. Мищенкова) входит более 1 млн словоупотреблений, доступных в поиске на лето 2019 г., и подготовлено пополнение в 200 тыс. словоупотреблений. В финско-русскую часть включены художественные, научные и публицистические тексты, в том числе 50 статей новостного агентства YLE. Хронологический охват корпуса — с 1880-х годов, включая финскую классическую прозу (А. Киви, Ю. Ахо, М. Лассила, Ф. Э. Силланпяя, М. Валтари, М. Ларни и др.) Русско-финская часть представлена пока только романом Солженицына «В круге первом», а также теми договорами в коллекции межгосударственных договоров с Финляндией, которые «диктовались» российской (советской) стороной и с большой долей вероятности были переведены с русского, хотя, как известно, определение направления перевода является существенной проблемой для нехудожественных текстов в целом, далеко не только юридических.

В дальнейшем планируется пополнить финский корпус текстами, уже входящими в разрабатываемые в университете Тампере корпуса ПарРус и ПарФин.

1.6. Чешский корпус

Чешский язык долгое время был заметной лакуной в параллельном корпусе НКРЯ — и это вызывало несомненные сожаления, учитывая как большое число славянских языков, уже представленных в двуязычных парах, так и существенное развитие корпусной лингвистики в Чехии, в том числе и в области параллельных корпусов. С 2018 года чешский корпус разрабатывается при участии Т. А. Малышевой (инициатора проекта), в том числе на базе текстов, взятых из параллельного корпуса А. Барентсена ASPAC. Как и в случае финского проекта, установлено сотрудничество (взаимовыгодный обмен текстами) с параллельным корпусом InterCorp в рамках Чешского национального корпуса (http://ucnk.korpus.cz/intercorp/). Для пословной разметки используется морфологический анализатор компании «Яндекс».

Объём корпуса уже в его пилотной версии, доступной с 2019 г., превосходит миллион словоупотреблений. В него входит несколько современных публицистических текстов, а также классические художественные произведения (четыре части гашековского «Швейка», «Невыносимая лёгкость бытия» М. Кундеры, Пушкин, Чехов, Булгаков, Ильф и Петров, Набоков и др.)

1.7. Шведский корпус

Шведский корпус, инициатором и основной разработчицей которого является Н. В. Перкова, как уже сказано, успел войти в пятерку крупнейших параллельных корпусов НКРЯ. Подобно ранее начавшемуся проекту Н. В. Перковой (латышскому корпусу), ставится цель хронологически репрезентативно охватить шведский «культурный канон» от Стриндберга до Бойе и Лагерквиста, представленный в русских переводах, а также наиболее важные произведения современной шведской прозы, довольно активно переводящейся на русский (Д. Ваттин, К.-Й. Вальгрен, К. Киери и др.). При этом не забывается и противоположное — русско-шведское направление, также представленное как классикой, так и современными романами (М. Шишкин, М. Степнова, Л. Петрушевская и др). На следующем этапе в корпус введены также современные новостные и аналитические тексты из шведской прессы; соответствующими работами под руководством К. Окерман-Саркисян и О. Янссон занимались студенты Упсальского университета М. Лундгрен и Э. Маттссон.

Тексты шведского корпуса размечены при помощи анализатора Stagger. Данный анализатор [Östling 2013] в современной его версии отличается высоким качеством морфологической разметки с автоматическим снятием омонимии, при разметке реализуется нейросетевая модель.

2. Новые задачи

2.1. Лингвоспецифичная разметка

На этапе, описанном в [Сичинава 2015], масштаб лингвоспецифичности разметки в параллельных корпусах НКРЯ был сравнительно невелик: конкретный набор тегов для грамматических категорий разных языков различался, однако схема морфологической разметки была одинаковой и заключалась в наборе частеречных и грамматических тегов (линейная упорядоченность этого набора при поиске роли не играет), аналогично принятому в НКРЯ с неснятой омонимией. В современном состоянии корпуса соблюдается принцип, согласно которому сопоставимые части речи и значения грамматических категорий в разных языках размечаются одинаково. В основном разметка ранее не встречавшихся граммем в новых языках (а инвентарь их вырос значительно: например, различные направительные и местные падежи в прибалтийско-финских, типы конвербов в бурятском и др.) следует «Лейпцигским правилам глоссирования» (https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf), хотя и с отклонениями, необходимыми для сохранения унифицированной разметки «со старыми» корпусами в составе НКРЯ (например, прошедшее время — латинизированное **praet**, а не англизированное **PST**).

Однако на настоящем этапе, как благодаря увеличению типологического разнообразия задействованных языков и письменностей, так и благодаря использованию

более сложных морфологических анализаторов, существенно расширился набор дополнительных параметров разметки, поиск по которым может быть релевантен для работающего с параллельным корпусом. Соответственно, разные языки имеют теперь не только разные наборы тегов (tagsets), но и разный набор атрибутов в пословной XML-разметке.

Наиболее масштабно отличается от среднекорпусного стандарта разметка китайских текстов. Это, разумеется, диктуется как «экзотической» (иероглифической) письменностью, требующей для неспециалиста дублирования в латинской транслитерации и хотя бы упрощенного глоссирования (ранее в корпусе такое решение, как и словарная разметка переводов лексики, уже принималось для армянского языка), так и типологическими отличиями китайского от ранее представленных в корпусе языков, в основном индоевропейских. Простой дихотомии «лексика vs. грамматика» в пословной разметке в целом ряде случаев недостаточно, кроме того, проблемным, в том числе в связи с письменностью, не знающей словоразделов, здесь является само словоделение. Размечаемые сегменты (токены), с точки зрения устройства XML-разметки аналогичные словам в других корпусах, в китайском корпусе делятся на «лексические» (содержащие один или несколько иероглифов) и «грамматические» (один иероглиф). «Грамматические» сегменты задаются списком (разработан Л. С. Холкиной), в большинстве случаев они могут быть омонимичны «лексическому» сегменту. «Грамматические» сегменты получают специфические грамматические пометы: «MOD» (модальная частица 了), «PFV» (перфектив), «PRG» (прогрессив), «PST» (прошедшее время), «EVAL» (оценка действия), «QUEST» (общий вопрос), «CAUS» (каузатив), «PL» (множественное число), «BA» (вынесение объекта в позицию перед глаголом), «ATRN» (определение к имени), «ATRV» (определение к глаголу), «PASS» (пассив), «DIR» (директивные частицы). Кроме того, часть «лексических» сегментов получают по словарю разметку одного грамматического (в данном случае словоклассифицирующего) параметра — классификатора, с которым сочетается данное слово, аналогично роду в европейских языках. Для разметки, как уже указано, используется китайско-английский словарь CEdict, из которого берутся все представленные в словаре толкования для того или иного лексического сегмента. Кроме того, китайским словам и предложениям приписана транслитерация в системе пиньинь, взятая из указанного словаря (с заменой цифрового обозначения тонов на диакритики) и снабженная «европеизированной» пунктуацией. Если один и тот же сегмент имеет несколько транслитераций, в том числе различающихся тонами, то приводятся через знак «/» все варианты (сохраняем автоматически введенное словоделение):

(1) Tiānqì shì nàyàng cháoshī **hé/hè/hú/huó/huò** duō wù, hǎobù róngyì cái tiānliàng. **cóng/cōng/zòng** chēxiāng chuāngkǒu wàng qù, tiělù zuǒyòu 10 bù lù **yuǎn/yuàn de/dī/dí/dì dìfāng/dìfang** jiù hěn nánkàn qīng shénme **dōngxī/dōngxi**.
Было так сыро и туманно, что насилу рассвело; в десяти шагах, вправо и влево от дороги, трудно было разглядеть хоть что-нибудь из окон вагона. [Достоевский, Идиот].

В начальной части «Идиота» на 46 671 китайский сегмент (аналог слова), выделенный программой словоделения, приходится 16 893 знака неоднозначности «/» — то есть в среднем один лишний транскрипционный вариант на три сегмента.

Таким образом, каждой китайско-русской выровненной паре предложений отвечает тройка строк — одна с русским текстом и две с китайским, в иероглифической и транслитерированной форме; кроме того, транслитерация сегментов продублирована в качестве дополнительного поля пословного разбора в иероглифической версии текста.

Разумеется, такая система имеет определенные недостатки, поскольку предусматривает маловероятные варианты разбора и транслитерации (а значительное количество альтернативных переводов и чтений характерно как раз для высокочастотных сегментов) и в то же время «жёсткий» выбор единственного варианта словоделения, однако она открыта для дальнейших улучшений и фильтрации маловероятных словарных разборов.

Переводная словарная информация (в поле sem) содержится также в бурятском корпусе, куда она введена Т. А. Архангельским на основании отсканированного и распознанного В. Ивановым словаря. Данная разметка покрывает большинство словоупотреблений, но нуждается в усовершенствовании (ручная чистка переводов, добавление отсутствующих лемм). Пример из «Капитанской дочки» Пушкина:

(2) <para id="17"><se lang="ru" variant_id="0">Я считался в отпуску до окончания наук.</se>
<se lang="bu" variant_id="1"><w><ana lex="зүгөөр" gr="A" sem="но, однако_же, хотя"></ana>Зүгөөр</w> <w><ana lex="би" gr="S,nom" sem="я"></ana>би</w> <w><ana lex="эрдэм" gr="S,nom" sem="наука"></ana>эрдэм</w> <w><ana lex="шудалха" gr="V,conv,simult1" sem="изучать, исследовать"></ana>шудалжа</w> <w>дүүргэтэрээ</w>, <w><ana lex="табилга" gr="S,dat" sem="разрешение"></ana>табилгада</w> <w><ana lex="гэхэ" gr="V,conv,simult1" sem="говорить"></ana>гэжэ</w> <w>тоотой</w> <w><ana lex="байха" gr="V,partcp,nonpast,1p,sg" sem="быть"></ana>байгааб</w>.</se>
</para>

Дополнительный унифицированный атрибут «словообразование» введен для эстонского (корпус этого языка уже существовал с 2015 г.) и финского корпусов. Как известно, в морфологии прибалтийско-финских языков большую роль играют продуктивные сложные слова, получающие в поле lex разбор как единая лексема, а в поле wordf — как сочетание нескольких основ (разделенных знаком плюса). Такая лексика получает дополнительную помету **compos**:

(3) Финский ‘мировая война’:
<w><ana lex="maailmansota" **wordf**="maa+ilman+sota" gr="S,**compos**,gen,sg" />maailmansodan</w>

(4) Эстонский ‘отечественная война’:
<w><ana lex="isamaasõda" **wordf**="isa+maa+sõda" gr="S,**compos**,gen,sg/>isa maasõja</w>

И наоборот, на данном этапе в параллельном корпусе уже разрешены орфографически неоднословные леммы и словоформы неизменяемых частей речи, содержащие пробел (в русском корпусе [Плунгян, Ляшевская, Сичинава 2005] они относились к «сложным лексическим единицам», обязательно сопровождаемым пословным разбором):

(5) Литовский ‘когда-нибудь’
<w><ana lex="kada nors" gr="ADV=pos"/>kada nors</w>

(6) Латышский ‘ведь’, ‘уже’
<w><ana lex="jau arī" gr="PART="/>jau arī</w>

2.2. Региональная вариативность языков

Параллельный корпус лишь в очень ограниченной мере может быть репрезентативным — его представительность ограничивается необходимым «оппортунистическим» критерием, то есть он в принципе может включать только тексты, которые уже были переведены, что искажает пропорции разных метатекстовых параметров по сравнению с генеральной совокупностью оригинальных текстов. Поэтому особую ценность представляют доступные в переводах тексты, представляющие те или иные региональные варианты литературного языка. Ранее в параллельных корпусах НКРЯ уже присутствовали образцы как британского, так и американского английского, а также тексты, созданные в конце XIX и начале XX в. на обоих исторических вариантах литературного украинского языка — западном (И. Франко, О. Кобылянская и др.) и центральном (И. Нечуй-Левицкий, М. Коцюбинский и др.). Сейчас же цель включения в корпус образцов полицентричных языков стала одной из важнейших.

Например, выросший почти вдвое за последние годы трудами К. О. Мищенковой испанский корпус, помимо образцов собственно кастильской прозы, по замыслу основной разработчицы, теперь включает в себя художественные и публицистические тексты, созданные авторами из следующих латиноамериканских стран: Аргентина (это не кто иной, как Че Гевара), Колумбия (Г. Гарсиа Маркес), Куба (А. Карпентьер), Мексика (К. Фуэнтес), Парагвай (А. Роа Бастос), Перу (М. Варгас Льоса; причем включенный в корпус его роман посвящен событиям в Доминиканской республике); а помимо них, двуязычный автор из Барселоны Э. Мендоса, пишущий также по-каталански. Тексты СМИ (переведенные на сайте inosmi.ru), пополнившие испанский корпус, происходят с испанских, аргентинских, венесуэльских, а также международных (имеющих локальную версию в нескольких испаноязычных странах) сайтов.

Финский корпус, также развиваемый К. О. Мищенковой, включает в себя тексты Ю. Конкка, ингерманландца, писавшего на стандартном финском, но испытавшем определенное влияние родного диалекта. В свою очередь, в шведский корпус, архитектурой художественной части которого занимается Н. В. Перкова, входят (или

запланированы к включению) произведения представителей «финляндского шведского»: Г. Парланда, Т. Янссон, М. Фагерхольм, Ч. Вестё, переводчика Я. Даля.

В перечисленных случаях региональная вариативность литературного языка значима в разной степени, а в некоторых из них, возможно, само ее наличие является исследовательской задачей, требующей особого доказательства, тем не менее разнообразие текстов по географическому признаку — важный параметр, который потенциально следует учитывать в лингвистическом исследовании.

Вот примеры латиноамериканизмов в текстах из испанского параллельного корпуса:

(7) Seguro que si estabas *vos* lo sacabas y Ana María creo que también, ya que no tienen esos complejos nochísticos que me dan a mí. [Ernesto Che Guevara. Notas de viaje (1952–1953)].
Уверен, что, будь *ты* на моем месте, ты бы его вытащила, и Ана Мария, думаю, тоже, поскольку у вас нет связанных с темнотой комплексов. [Эрнесто Че Гевара. Дневник мотоциклиста (А. Ведюшкин, 2014)].

(8) «No podía dormir por la rabia de estar pensando en él, pero lo que más rabia me daba era que *mientras más* rabia sentía, más pensaba». [Gabriel García Márquez. Vivir para contarla (2002)].
Я не могла спать от бешенства, что он не идет у меня из головы, но *чем больше* я бесилась, тем больше о нем думала. [Габриэль Гарсиа Маркес. Жить, чтобы рассказывать о жизни (С. Марков, Е. Маркова, А. Малоземова, В. Федотова, 2012)].

(9) Pero, para hacer frente a sus frugales necesidades, en sus horas libres fue vendedora en un supermercado, *mesera* en una pizzería de Boston…[Mario Vargas Llosa. La Fiesta del Chivo (2000)].
Однако на жизнь, хотя потребности у нее были более чем умеренные, ей пришлось зарабатывать в свободные часы продавщицей в супермаркете, *официанткой* в бостонской пиццерии… [Марио Варгас Льоса. Нечестивец, или Праздник Козла (Людмила Синянская, 2004)].

(10 Los comedores donde se hartaban, las aguas infectadas, los *excusados* públicos y apestosos y las recámaras donde ellos cogían y roncaban… [Carlos Fuentes. Gringo viejo (1985)].
Я сжег грязные столовые, зараженные водоемы, зловонные *отхожие места*; конуры, где бесились и рычали псы… [Карлос Фуэнтес. Старый гринго (Маргарита Былинкина, 2010)].

2.3. Жанровое разнообразие корпусов

Жанровое разнообразие корпусов, в 2015 году лишь намечавшееся, в 2019 году является одной из главных целей, причем эта цель учитывается с самого начала создания новых языковых пар. Так, новый башкирский корпус — единственный, где

нехудожественные тексты преобладают над художественными, а также, как указано выше, включены тексты из Википедии и словарей.

Активно используется такой источник нехудожественных текстов, как русские переводы иностранной прессы на сайте ИноСМИ.ру. Несмотря на наличие только одного направления перевода, трудности в автоматизации скачивания, не всегда доступные тексты оригинальных СМИ (часто требующие отдельной подписки на платные версии их сайтов) и нередко вольный перевод, эту коллекцию переводной публицистики на десятках языков из сотен изданий трудно переоценить. В частности, тексты оттуда пополнили испанский, итальянский, шведский, английский, чешский, финский корпуса; готовится значительное пополнение и армянского корпуса.

Научные и философские тексты включаются во французский корпус — в частности, туда вошли работы Ж. Женетта, Ж. Кокто, М.-П. Рей, М. Бахтина, Н. Бердяева. Заметный прирост таких текстов произошёл и в итальянском корпусе (историческая работа Дж. Боффы «От СССР к России», физический трактат Т. Редже «Этюды о вселенной», «Кризис западной философии» В. Соловьева, а также «Проблемы поэтики Достоевского» Бахтина, причем перевод сделан с несколько иной редакции оригинала, чем во французском корпусе).

В испанский корпус введена подборка двуязычных путеводителей по Испании (подготовленная и выровненная Т. Горожанкиной) — данный тип дискурса в параллельном корпусе ранее был не представлен (да даже и в одноязычном НКРЯ он учитывается лишь в очень ограниченной степени). В китайском корпусе имеются деловая переписка и переводы Евангелия, делавшиеся с учетом славянских текстов (ранее библейские тексты включались в белорусский корпус НКРЯ). В итальянский корпус включены пресс-коммюнике и материалы российско-итальянской торговой палаты.

Жанровое разнообразие способствует разнообразию лексики и конструкций в корпусе: например, русское сочетание (грамматикализирующееся как предлог) *в рамках* в итальянском корпусе встретилось 24 раза, причем только 3 раза в художественных текстах; а в армянском корпусе, где нехудожественных текстов пока нет — только 1 раз (как канцеляризм в прямой речи у Толстого).

2.4. Развитие поливариантных двуязычных корпусов

Напомним, что в рамках особого проекта, начавшегося в 2012 г., французский корпус был в порядке эксперимента пополнен поливариантными переводами («Шинель», «Нос», «Обломов»). Основной целью этого проекта (продолжением его стал целый ряд проектов РНФ, РФФИ, частного фонда «Династия» и Швейцарского научного фонда, выполняемых в ИПИ РАН под руководством Анны А. Зализняк и О. Ю. Иньковой) стало создание надкорпусных баз данных по видо-временным формам, коннекторам и дискусивным словам. Позже в эти базы была включена и соответствующим образом размечена «Шинель» Гоголя, представленная в 15 итальянских переводах, и другие итальянские тексты с одним вариантом

перевода. Надкорпусные базы данных и соответствующие публикации доступны на сайтах http://a179.frccsc.ru/PublicLingvoProjects/main.aspx и http://a179.frccsc.ru/RFH41002/main.aspx.

Поливариантные корпуса ценны для выявления внутриязыкового пространства возможностей и точек варьирования. Нередко переводчики, будучи знакомы с предыдущими переводами, сознательно от них отталкиваются. Эта тенденция продолжается, как в «старых», так и в «новых» языковых разделах. Например, в латышском корпусе «Четыре поездки» В. Лациса представлены в переводах Г. Цейтлина и В. Ругайса, рассказы Чехова — в различных латышских переводах П. Калвы, Р. Эзеры, А. Гревини, О. Калнциемса, А. Курцийса; в шведском корпусе «Шляпа волшебника» Т. Янссон — в переводах В. Смирнова и Л. Брауде, а «Пеппи/Пиппи Длинныйчулок» — в переводах Л. Лунгиной и Л. Брауде, в немецком корпусе «Капитанская дочка» Пушкина — в переводах Ф. Отто и Ф. Фриш (?). Во французском корпусе, развиваемом в ходе упомянутых выше проектов ИПИ РАН, дважды представлено сочинение «Difficulté d'être» Ж. Кокто — в версиях под названиями «Тяжесть бытия» (Л. М. Цывьян) и «Трудность бытия» (М. Л. Аннинская)[2]. Поливариантными являются и файлы с различными китайскими переводами Евангелия (см. выше). Пока не все эти альтернативные версии перевода сведены в единые файлы, однако могут искаться в Корпусе одновременно и сопоставляться.

2.5. Развитие функциональности поиска

На текущей платформе Яндекс.Поиска ряд размеченных параметров пока не реализован (например, типология неточности перевода, о которой см. Сичинава 2015, или словообразование сложных слов). Частично подобный поиск может быть осуществлен в разрабатываемых ИПИ РАН надкорпусных базах данных по французскому и итальянскому корпусам (см. выше, 2.4).

Поставлена задача внедрить эти параметры в поиск после переезда корпуса на новые технологические рельсы. В ожидании же этого момента некоторые функции реализованы на движке «Цакорпус» (доступен по адресу https://bitbucket.org/tsakorpus/tsakonian_corpus_platform/src/default/; корпуса публикуются на сайте Linghub.ru) Т. А. Архангельским. В частности, возможен одновременный поиск в двух языках (например, можно выяснить, когда шведское fika переводится на русский как [*выпить*] *кофе*, а когда русское *простор* — как шведское vidd), получение частотных списков слов, статистика сочетаемости, статистика по дате создания текста.

[2] А в чешском корпусе одновременно (по случайному совпадению) появилась и «Nesnesitelná lehkost bytí» («Невыносимая лёгкость бытия») М. Кундеры; правда, существует как будто бы только один перевод этого романа.

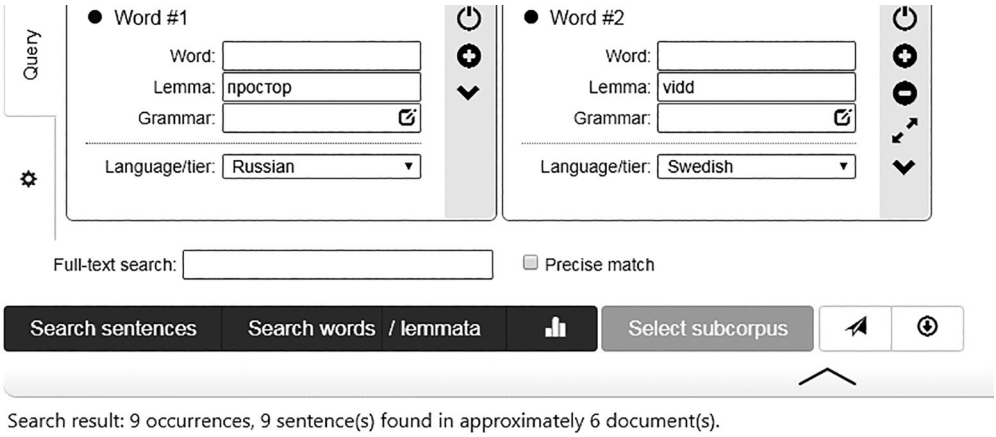

Рис. 1. Образец поискового интерфейса на linghub (движок «Цакорпус») для одновременного поиска в обоих языках

2.6. Расширение многоязычных корпусов

В рамках проекта Йенского университета (руководители Р. фон Вальденфельс и И. С. Левин) ведется работа по созданию многоязычного параллельного корпуса ParaPooh, который включает в себя переводы «Винни-Пуха» на более чем 50 языков из коллекции автора настоящей статьи (некоторые переводы приобретены специально для корпуса). Сканированием и распознаванием текстов занимается О. Д. Шиншинова (МГУ). По окончании работ тексты пополнят как многоязычный корпус ParaSOL [von Waldenfels 2006] так и, по крайней мере частично, НКРЯ (в котором в настоящее время «Винни-Пух» представлен на 16 языках).

Анализ многоязычного текста «Винни-Пуха» по материалам расширенной коллекции позволяет построить следующую сеть NeighbourNet (о построении сетей расстояний между данными и их использовании в кластерном анализе см. Сичинава 2015) для граммемы плюсквамперфекта в 24 идиомах Европы.

Всего в тексте представлено 258 предложений, в которых плюсквамперфект отмечен хотя бы в одном языке. В 89 из этих

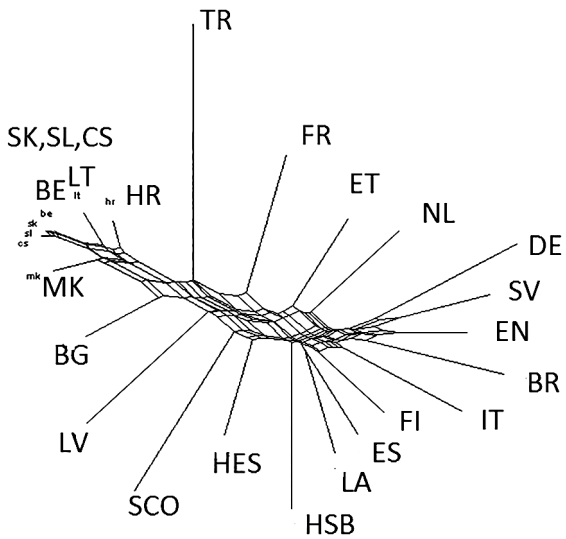

Рис. 2. Сеть NeighbourNet для плюсквамперфекта в переводах «Винни-Пуха»

случаев плюсквамперфекта нет в английском оригинале; часто это связано с использованием нефинитных средств выражения (перфектного инфинитива, номинализации, герундия). Однако выделяется, в частности, класс случаев «ошибочное решение / исправление»; для его кодирования в английском используется претерит, а некоторые языки выбирают плюсквамперфект, маркирующий отмененный результат или смежное значение:

(11) англ. I *forgot*.
швед. Det *hade* jag *glömt*.
латыш. Es *biju aizmirsis.*
верхнелуж. *Běch* to *zabył.*
'Я позабыл [что сам съел мёд]'.

Интересен случай *ирреального* действия, предшествовавшего реальному; здесь, например, в близкородственном нидерландском языке также выбирается плюсквамперфект, в то время как в оригинале выступает простое прошедшее:

(12) англ. You gave him don't you remember — a little — a little — I gave him a box of paints to paint things with. — That was it. — Why *didn't* I give it to him in the morning?
нид. Jij hebt hem een... Weet je het niet meer?... een kleine... een kleine... — O, ja...! Een verfdoosje om mee te verven! — Precies. — Maar waarom *had* ik het hem niet meteen 's morgens *gegeven*?

'Ты подарил ему, разве ты не помнишь, ну эту... маленькую... маленькую... — Я подарил ему коробочку с красками, чтобы он мог ими рисовать. — Именно так. — А почему я не *подарил* ее ему тем утром?'

Приведенный выше граф NeighbourNet показывает, что языки Европы по употребительности этой формы в различных контекстах распределены между двумя полюсами:

максимально частое употребление, квазиобязательное использование в таксисных контекстах: германские языки — английский [EN, язык оригинала], нидерландский [NL], немецкий [DE], шведский [SV] (их кластеризация на графе, несмотря на указанные выше частные различия, демонстрирует схождение генетических и типологических характеристик); бретонский [BR, кельтские]; латинский [LA, италийские] и два романских языка — испанский [ES], итальянский [IT]; прибалтийско-финские языки — финский [FI] и эстонский [ET], которые, однако, не образуют кластера друг с другом);

редкое использование; в контрфактических или антирезультативных контекстах (словацкий [SK], словенский [SL], чешский [CS], белорусский [BE]); несколько более частое применение для подчеркивания последовательности временных планов (македонский [MK], болгарский [BG], хорватский [HR], литовский [LT][3]).

[3] В целом плюсквамперфект используется в литовском языке чаще, чем в славянских. Возможно, это индивидуальная особенность данного перевода. Кроме того, надо учесть, что английский

Об употреблении плюсквамперфекта в славянских языках (с примерами близких к «инвариантным» контекстов) подробнее см. [Сичинава 2019].

Промежуточную позицию между этими двумя полюсами занимают турецкий [TR] (употребления соответствующих форм в этом языке вообще стоят особняком; они маркируют дигрессии, состояния во временном плане прошлого, недостигнутый результат), латышский [LV], французский [FR], верхнелужицкий [HSB]. Отметим, что литературные микроязыки на диалектной базе (скотс [SCO] и гессенский [HES]) демонстрируют заметно более редкое употребление плюсквамперфекта, чем стандартные английский и немецкий соответственно.

3. Промежуточные итоги

Дальнейший план развития параллельных корпусов, после стабилизации новой версии Яндекс.Поиска для НКРЯ, включает в себя:

а) дальнейшее добавление новых языковых пар с русским (предварительно речь идёт о словенском, сербском, словацком, венгерском, японском, португальском, а, возможно, и ряде других языков);

б) развитие существующих языковых пар (например, добавление болгарско-русских текстов — сейчас в корпусе представлены только русско-болгарские; добавление нехудожественных текстов в корпуса, где они сейчас отсутствуют, например, в армянский или литовский);

в) полная реализация поиска по всем размеченным параметрам, в том числе по словообразованию и неточностям перевода;

г) подключение средств визуализации, например, построения графиков, аналогично существующим в основном и газетном корпусах НКРЯ.

Литература

Михайлов М.Н., Хярме Ю. Параллельные корпуса художественных текстов в Тамперском университете // Русский язык за рубежом. Финская русистика (специальный выпуск). М., 2015. С. 16–19.

Морозова В.А. Построение русско-китайского параллельного корпуса текстов: реализация нейросетевой модели для реализации автоматического словоделения иероглифических текстов. Курсовая работа. ВШЭ, 2018.

Орехов Б.В. Проблемы морфологической разметки башкирских текстов // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. Казань: Изд-во «Фэн» Академии наук РТ, 2014. С. 135–140

оригинал мог повлиять на выбор времени весьма ограниченно, поскольку перевод В. Чепайтиса первоначально делался с польского текста И. Тувим и лишь потом правился по оригиналу. Как видим, у этого в принципе полезного «нивелирующего» свойства есть и обратная сторона (демонстрирующая, между прочем, и ограниченность возможностей корпусного исследования для граммем, активно конкурирующих с синонимичными).

Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М., 2005. С. 111–135.

Сичинава Д. В. Параллельные тексты в составе Национального корпуса русского языка: новые направления развития и результаты // Труды Института русского языка РАН, 2015. № 6. С. 194–235.

Сичинава Д. В. Славянский плюсквамперфект: пространства возможностей // Вопросы языкознания. 2019, № 1. С. 30–57.

Шведова М., фон Вальденфельс Р., Яригін С., Крук М., Рисін А., Возняк М. Генеральний регіонально анотований корпус української мови (ГРАК). Київ, Осло, Єна, 2017–2019. [Электронный ресурс]. URL: uacorpus.org.

Якубов М. Н. Построение китайско-русского корпуса параллельных текстов: проблемы выравнивания, словоделения и лексико-семантической разметки. Выпускная квалификационная работа. ВШЭ, 2017.

Dahl, Ö., Koptjevskaja-Tamm, M. (eds.) Circum-Baltic languages. Typology and contact. Vol. 1-2, Amsterdam—Philadelphia: Benjamins, 2001.

Östling, R. Stagger: an Open-Source Part of Speech Tagger for Swedish. // Northern European Journal of Language Technology, 2013, Vol. 3, Article 1, pp 1–18.

Perkova, N., Sitchinava, D. On the Development of a Latvian-Russian Parallel Corpus // Skadiņa, I., Rozis, R. (eds.). Human Language Technologies — The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016, pp. 130–135. IOS Press, Amsterdam, 2016.

Rimkutė, E., Daudaravičius, V., Utka, A. Morphological Annotation of the Lithuanian Corpus. 45th Annual Meeting of the Association for Computational Linguistics. // Workshop Balto-Slavonic Natural Language Processing, 2007, pp. 94–99.

Sitchinava D., Perkova, N. Bilingual Parallel Corpora Featuring the Circum-Baltic Languages within the Russian National Corpus. Digital Humanities in the Nordic Countries Proceedings of the Digital Humanities in the Nordic Countries 4th Conference Copenhagen, Denmark, March 5-8, 2019, pp. 495–502.

von Waldenfels R. Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment // Beitrage der Europäischen Slavistischen Linguistik (POLYSLAV) 9. München, 2006, S. 123–138.

Dmitri V. Sitchinava
Vinogradov Russian Language Institute of the Russian Academy of Sciences
National Research University Higher School of Economics
(Russia, Moscow)
mitrius@gmail.com

ON PARALLEL TEXTS WITHIN THE RUSSIAN NATIONAL CORPUS : NEW LANGUAGES AND NEW CHALLENGES

The paper discusses the main trends in the development of the parallel corpora within the RNC since 2015. The New languages section deals with seven new language pairs that emerged during this period, their architecture and tagging.

Compared with the list of languages that form bilingual parallel pairs with Russian available in 2015, the following new languages have appeared in the RNC: Bashkir, Buryat, Chinese, Czech, Finnish, Lithuanian, Swedish. Creating the parallel subcorpora comes as the combined efforts of autonomous Russian and foreign teams coordinated by the team of developers of the RNC in Moscow. Virtually all the new languages offer specific challenges for the Corpus developers with regard to their annotation. The size of some of the language pairs already available in 2015 has also significantly increased in four years.

The New challenges section goes further to explore some general trends of parallel corpora across different language pairs such as regional/national language varieties, representativeness with regard to text genres, new annotation and search types etc. At the present stage, both due to an increase in the typological diversity of the languages and scripts involved, and through the use of more complex morphological analyzers, the set of additional markup parameters has been significantly expanded. The purpose of incorporating polycentric language samples into the corpus has become one of the most important. The genre representativeness of corpora, which was only being planned in 2015, is one of the main goals in 2019, and this goal is taken into account from the very beginning of the creation of new language pairs.

A case study is dedicated to studying pluperfect in 24 languages of Europe in an expanded multilingual corpus. The analysis of a multilingual text based on the material of an extended collection allows us to construct a network of distances between the data for the Pluperfect category in 24 European languages/lects.

Key words: parallel corpora; bilingual corpora; multilingual corpora; annotation; representativeness; pluperfect

References

Dahl, Ö., Koptjevskaja-Tamm, M. (eds.) *Circum-Baltic languages. Typology and contact*. Vol. 1-2, Amsterdam—Philadelphia: Benjamins, 2001.

Lyashevskaya O.N., Plungian V.A., Sitchinava D.V. [On morphological standard of the Russian National Corpus] // *Natsional'nyi korpus russkogo yazyka*: *2003-2005*. *Rezul'taty i perspektivy* [The Russian National Corpus: 2003-2005. Results and prospects]. Moscow, 2005, pp. 111–135. (In Russ.)

Mikhailov M. N., Härme J. [Parallel Corpora of Fiction at the University of Tampere] // *Russkii yazyk za rubezhom. Finskaya rusistika* (spetsial'nyi vypusk) [Russian language abroad. Finnish Russian studies (special issue)]. Moscow, 2015, pp. 16–19. (In Russ.)

Morozova V. A. *Postroenie russko-kitaiskogo parallel'nogo korpusa tekstov: realizatsiya neirosetevoi modeli dlya realizatsii avtomaticheskogo slovodeleniya ieroglificheskikh tekstov*. Kursovaya rabota. [Building Russian-Chinese parallel corpus : applying of neural network model to word segmentation of hieroglyphic texts] NRU HSE, 2018. (In Russ.)

Östling, R. Stagger: an Open-Source Part of Speech Tagger for Swedish // *Northern European Journal of Language Technology*, 2013, Vol. 3, Article 1, pp 1–18.

Orekhov B. V. [Problems of morphological marking of Bashkir texts]. *Trudy Kazanskoi shkoly po komp'yuternoi i kognitivnoi lingvistike TEL-2014* [Proceedings of the Kazan summer school in computer and cognitive linguistics TEL-2014]. Kazan', "Fen" Publ., 2014, pp. 135–140. (In Russ.)

Perkova N., Sitchinava D. On the Development of a Latvian-Russian Parallel Corpus // Skadiņa, I., Rozis, R. (eds.). *Human Language Technologies — The Baltic Perspective*: Proceedings of the Seventh International Conference Baltic HLT 2016, pp. 130–135. IOS Press, Amsterdam, 2016.

Rimkutė E., Daudaravičius V., Utka A. Morphological Annotation of the Lithuanian Corpus. 45th Annual Meeting of the Association for Computational Linguistics. *Workshop Balto-Slavonic Natural Language Processing*, 2007, pp. 94–99.

Sitchinava D., Perkova N. Bilingual Parallel Corpora Featuring the Circum-Baltic Languages within the Russian National Corpus. *Digital Humanities in the Nordic Countries*. Proceedings of the Digital Humanities in the Nordic Countries 4th Conference Copenhagen, Denmark, March 5-8, 2019, pp. 495–502.

Sichinava D. V. [Parallel texts within the Russian National Corpus: new trends of developments and new results]. *Trudy Instituta russkogo yazyka RAN* [Proceedings of the V. V. Vinogradov Russian Language Institute], 2015, no. 6, pp. 194–235. (In Russ.)

Sichinava D. V. [Slavic Pluperfect: foci of variation] *Voprosy jazykoznanija* [Topics in the study of language]. 2019, no. 1, pp. 30–57. (In Russ.)

Shvedova M., von Waldenfels R., Jaryhin S., Kruk M., Rysin A., Woźniak M. *Heneral'nii rehional'no anotovanyi korpus ukrajins'koji movy (HRAK)* [General Regionally Annotated Corpus of Ukrainian (GRAC)]. Kyiv, Oslo, Jena, 2017–2019. Available at: uacorpus.org.

von Waldenfels R. Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment. *Beitrage der Europäischen Slavistischen Linguistik (POLYSLAV)* 9. München, 2006, pp. 123–138.

Yakubov M. N. *Postroenie kitaisko-russkogo korpusa parallel'nykh tekstov: problemy vyravnivaniya, slovodeleniya i leksiko-semanticheskoi razmetki.* Vypusknaya kvalifikatsionnaya rabota. [Building Chinese-Russian parallel corpus : alignment, word segmentation, lexical and semantic annotation]. NRU HSE, 2017. (In Russ.)