Д. В. Сичинава
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
mitrius@gmail.com

ПАРАЛЛЕЛЬНЫЕ ТЕКСТЫ В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: НОВЫЕ НАПРАВЛЕНИЯ РАЗВИТИЯ И РЕЗУЛЬТАТЫ[1]

В статье представлено текущее состояние параллельных корпусов в составе НКРЯ и работа, проведенная над этими корпусами в период 2009–2015 гг. Параллельные корпуса НКРЯ включают следующие параллельные двуязычные (с русским) корпуса: английский, армянский, белорусский, болгарский, испанский, итальянский, латышский, немецкий, польский, украинский, французский, эстонский. Практически для всех из этих языков представлены обе симметричные языковые пары. В соответствии с мировым опытом построения параллельных корпусов, теперь входит многоязычный корпус, состоящий из 9 текстов и задействующий более 20 языков (в основном славянских). Как поливариантный с 2012 г. разрабатывается русско-французский корпус, включающий до 4 вариантов перевода для некоторых текстов. Поливариантные тексты включены и в многоязычный корпус. В параллельные корпуса теперь включаются как художественные, так и нехудожественные тексты с той же классификацией, что и в основном русском корпусе (публицистика, производственно-технические, учебно-научные, церковные, юридические тексты). Разработан инструментарий разметки несоответствий и вольностей при переводе (пропуск части предложения,

[1] Работа выполнена в рамках Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» и при поддержке РГНФ (грант № 15-04-12018).

вставка, значимая неадекватная замена, выполненная переводчиком), значительная часть текстов размечена с учетом этого инструментария. Тексты на большинстве языков получают морфологическую разметку. Рассмотрены конкретные примеры использования корпусов для исследования лексики и грамматики.

Ключевые слова: Параллельный корпус, разметка, многоязычный корпус, лексическая типология, грамматическая типология, лингвоспецифичная лексика, перфект

За время, прошедшее с выхода сборника «Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы», параллельные корпуса в составе НКРЯ (http://ruscorpora.ru/search-para.html) пережили очень важный этап развития как с технологической, так и с содержательной стороны. Теперь это один из самых объёмных (более 70 миллионов словоупотреблений) и важных проектов НКРЯ, задействующий несколько команд специалистов из разных стран (ср. публикацию [Sitchinava 2012]). Это направление получило поддержку Программы Президиума РАН «Корпусная лингвистика» (проект «Создание и развитие параллельных русско-иноязычных корпусов в Национальном корпусе русского языка», руководитель Д. О. Добровольский). Ряд параллельных корпусов (например, русско-французский мультиязычный, белорусский, польский) развивался в эти годы и как самостоятельные подпроекты, поддержанные отдельными грантами. Параллельные корпуса — едва ли не самый «многолюдный» проект в составе Национального корпуса; как и диктует специфика задачи, его команда — это целый ряд групп разработчиков, представляющих разные страны.

Назовём **языковой парой L1— L2** параллельный (под)корпус, состоящий из текстов, созданных в оригинале на языке L1 и переведённых на язык L2. В дальнейшем мы для обозначения двух симметричных пар «русский-L корпус» и «L-русский корпус» будем также использовать более краткое выражение **L-й параллельный корпус** (например, «армянский параллельный корпус» = «армянско-русская и русско-армянская языковые пары»).

В 2008 году Национальный корпус русского языка включал только три языковые пары: англо-русский, русско-английский и немецко-русский корпус. Эти корпуса включали в себя только художественные тексты.

За прошедшие шесть лет параллельные корпуса получили развитие в следующих направлениях:

1. Новые языковые пары. Сейчас параллельные корпуса включают следующие параллельные двуязычные (с русским) корпуса: английский, армянский, белорусский, болгарский, испанский, итальянский, латышский, немецкий, польский, украинский, французский, эстонский. Практически для всех из этих языков представлены обе симметричные пары, например, армянско-русский и русско-армянский (под)корпуса. В настоящее время нет болгарско-русских текстов (в НКРЯ включен готовый русско-болгарский корпус, составленный в Тырновском университете). Русско-немецкая языковая пара, отсутствовавшая в 2008 году, появилась и активно развивается. Часть из этих корпусов представляет собой небольшие по объёму пилотные проекты из нескольких текстов (особенно испанский и в какой-то мере латышский), часть — вполне полноценные корпуса (крупнейшими по объёму, помимо английского и немецкого, являются также украинский и польский). Существуют проекты создания и других двуязычных пар (например, венгерского, литовского, шведского корпусов).

2. Многоязычный (мультиязычный) корпус. В НКРЯ, в соответствии с мировым опытом построения параллельных корпусов, теперь входит многоязычный корпус, состоящий из 9 текстов и задействующий более 20 языков (в основном славянских).

3. Поливариантный корпус, включающий более одного перевода с языка L1 на язык L2. Как поливариантный с 2012 г. разрабатывается русско-французский корпус, включающий до 4 вариантов перевода для некоторых текстов. Поливариантные тексты включены и в многоязычный корпус. Имеются также русско-итальянские поливариантные тексты. Есть планы аналогичного развития также иных языковых пар (английские, немецкие тексты).

4. Жанровое разнообразие. В параллельные корпуса теперь включаются как художественные, так и нехудожественные тексты на основе той же классификации, что и в основном русском корпусе (публицистика, производственно-технические, учебно-научные, церковные, юридические тексты).

5. Средства разметки и выравнивания. Разработан инструментарий разметки несоответствий и вольностей при переводе (пропуск части предложения, вставка, значимая неадекватная за-

мена, выполненная переводчиком), значительная часть текстов размечена с учетом этого инструментария. Тексты на большинстве языков получают морфологическую разметку. Польские тексты размечены с автоматическим снятием омонимии, армянские — с рядом облегчающих поиск особенностей (перевод лемм на английский язык, версия в латинице). Для всех русских текстов в составе параллельных корпусов доступен такой же семантический поиск, что и в основном русском корпусе. Расширена и развита метаразметка текстов: сейчас метаинформация параллельного корпуса включает сведения о сфере функционирования текста, дате создания перевода, предусматривает указания названия текста и имени автора на нескольких языках. Разработана специальная программа-оболочка «Евклид» (автор Т. А. Архангельский), в которую включено автоматическое алгоритмическое выравнивание, осуществляемое программой HunAlign; в программе предусмотрена удобная корректировка ошибок автоматического выравнивания, разметка неточности в переводе (ср. выше), а также модуль заполнения таблицы метаразметки и конвертация текста в XML в формате UTF-8. При помощи этой программы один разметчик может самостоятельно собирать тексты, выравнивать и метаразмечать их, получая на выходе файлы XML и метатаблицы в формате, готовом для размещения в параллельном корпусе.

Ниже рассмотрим все эти пять направлений развития параллельных корпусов НКРЯ. Кроме того, будут рассмотрены конкретные примеры использования корпусов для исследования лексики (раздел 6) и грамматики (раздел 7).

1. Новые языковые пары и развитие старых

1.1. Английский корпус

Английский и немецкий параллельный корпуса, как и на момент публикаций в сборнике НКРЯ 2006–2008, разрабатываются под руководством А. А. Кретова и Д. О. Добровольского сотрудниками Воронежского университета (следует назвать вносящих особо активный вклад И. А. Меркулову, Ю. П. Плешкову, Е. Н. Подтележникову, Ю. А. Суворову, Кс. М. Шилихину).

В 2008–2014 гг. в английский параллельный корпус ежегодно включались новые художественные тексты писателей XIX — XXI вв., в том числе т. н. «классические» переводы, представляющие большое культурно-историческое значение, сопоставимое с оригинальными текстами на данном языке (например, переводы русских классиков, выполненные К. Гарнет, или «Над пропастью во ржи» Сэлинджера в переводе Р. Райт). Кроме того, в корпус введены предоставленные компанией ABBYY (проект ABBYY Lingvo) переведённые с английского нехудожественные тексты: технические (инструкции, документы и официальные отчеты организаций), учебно-научные, публицистические. В 2015 г. в корпус также включено представительное собрание сочинений В. В. Набокова как русского, так и американского периода, в оригиналах и переводах соответственно на английский и русский языки, в том числе авторских и авторизованных (тексты собраны и выровнены Я. А. Жеребцовой). Английский корпус остаётся самым большим из параллельных корпусов НКРЯ (18,1 млн словоупотреблений в англо-русской и 6,5 млн в русско-английской части), на его материале активно проводятся исследования лексики и грамматики (см. разделы 6 и 7), в том числе диахронические.

1.2. Армянский корпус

Создан армянский паралельный корпус размером 2,1 млн словоупотреблений, большая часть которых приходится на русско-армянские тексты. В основу корпуса положены переводы с русского языка на армянский (в основном классической художественной литературы), входящие в Восточноармянский национальный корпус (EANC; http://eanc.net). При отборе текстов обращалось внимание на репрезентативность с точки зрения включения максимального количества разных авторов и переводчиков. В корпус вошли оригиналы и переводы на армянский язык прозаических произведений М. Булгакова, И. Бунина, Н. Гоголя, И. Гончарова, Ф. Достоевского, В. Ерофеева, А. Куприна, О. Мандельштама, А. Н. Островского, Н. А. Островского, А. Пушкина, А. Рыбакова, М. Салтыкова-Щедрина, Л. Толстого, И. Тургенева, А. Фадеева, А. Чехова. В армянско-русскую языковую пару включены оригиналы и переводы на русский язык произведений Х. Абовяна, А. Бакунца и Д. Демирчяна.

Все армянские тексты снабжены морфологической разметкой с неснятой омонимией, основанной на морфологическом модуле EANC. Эта разметка включает в себя переводы армянских лемм на английский язык. Кроме того, учитывая особенности армянской графики, труднодоступной неспециалисту, все армянские тексты продублированы в латинской транслитерации (включая морфологическую разметку лемм). Три слоя — русский, армянский в графике и армянский в латинице — автоматически выровнены предложение в предложение.

Выравниванием текстов занималась Т. О. Шаврина, технической поддержкой нестандартной конвертации разметки и транслитерации — Т. А. Архангельский.

1.3. Белорусский корпус

Белорусско-русская языковая пара — в своей основе плод отдельного совместного проекта РГНФ и БРФФИ (№ 11–24–01004а/Bel «Корпусные сопоставительные исследования русского и белорусского языков и разработка параллельных электронных корпусов»). Корпус пополняется и после окончания этого гранта. Беларусь — страна, где в последние годы активно развиваются корпусные проекты, в том числе и с участием российских специалистов; в этом ряду параллельный корпус стал одним из первых (по крайней мере из доступных в Интернете). Руководство проектом с белорусской стороны взял на себя один из главных энтузиастов корпусной лингвистики в этой стране В. А. Кощенко (Национальная академия наук Беларуси), в разработке корпуса на разных этапах активно участвовали также И. В. Глинник, А. В. Зубов, И. Л. Копылов, О. В. Мицкевич, Е. Н. Скопинова, Ю. А. Стасевич (Коровко).

Именно в ходе работы над этим корпусом по предложениям работавших над ним белорусских специалистов была создана и существенно усовершенствована программа выравнивания «Евклид», появилась технология разметки неточностей в переводе (связанная со спецификой художественного перевода с «языков народов СССР»).

Белорусский корпус включает новостные тексты (материалы двуязычных новостных агентств «БелаПАН» и «Телеграф»), а также тексты существующих в двух языковых версиях белорусских

законов 1990–2000-х гг. Язык оригинала для этих текстов указан как белорусский условно, поскольку реальное направление перевода не всегда можно установить. Русско-белорусский компонент также недавно пополнился новым переводом библейских новозаветных книг, выполняемых библейской комиссией Белорусского экзархата РПЦ. В качестве оригинала с долей условности указан русский синодальный перевод Библии, на решения которого белорусский перевод в известной степени ориентируется, хотя делается с греческого оригинала. Текст белорусских новозаветных книг, включенных в корпус, частично снабжён знаками ударения.

В основу морфологической разметки белорусского корпуса положен модуль морфологии, разработанный И. В. Совпелем и усовершенствованный компанией «Яндекс» (в частности, этот модуль теперь частично обрабатывает так называемую тарашкевицу, или белорусское классическое правописание, с передачей предконсонантной мягкости: *сьвет, прывітаньне*). Ведется работа над усовершествованием этой разметки (в частности, включением в словарь таких слов, как, например, местоимений *іхні* и *ягоны*, частотных в текстах, но ограниченно представленных в словарных стандартах, ориентированных на советские нормы).

Общий объём белорусско-русского параллельного корпуса — 4,9 млн словоупотреблений в белорусско-русской и 1,8 млн в русско-белорусской части.

1.4. Болгарский корпус

Русско-болгарский компонент основывается на текстах, любезно предоставленных Великотырновским университетом (группой под руководством проф. Гочо Гочева, http://rbcorpus.com/ekip_rus.php). Тексты совокупным объёмом 3,6 млн словоупотреблений представляют собой оригиналы и переводы на болгарский язык произведений XX — XXI вв., причём включают не только художественные тексты, но и публицистику (например, фельетоны М. Кольцова). Все тексты в рамках российского проекта переконвертированы в формат XML и снабжены морфологической разметкой с неснятой омонимией на основе болгарского морфологического модуля компании «Яндекс».

1.5. Испанский и итальянский корпуса

В пилотной стадии в настоящее время находятся новые романские корпуса НКРЯ. В испанский и итальянский корпус входят как переводы с русского (например, «Анна Каренина» Толстого, «Град обреченный» Стругацких), так и переводы на русский (например, «Улей» Камило Хосе Селы, «Имя розы» Умберто Эко). Все эти тексты относятся к художественной литературе, задача жанровой репрезентативности пока не ставилась. Тексты получают морфологическую разметку на базе анализаторов компании «Яндекс», которая ещё нуждается в дополнительном усовершенствовании.

Над выравниванием испанских текстов работали В. С. Люсина и С. Ю. Бочавер, существует также проект сотрудничества в этом направлении с университетом Гранады (Р. Гусман Тирадо). Продолжается работа над развитием итальянского корпуса с участием Католического университета Милана (А. Бонола и В. Ноздеа) и Болонского университета (Ф. Бьяджини, Кс. Д. Балакина), в частности, корпус включает теперь поливариантные итальянско-русские тексты. К моменту выхода этой статьи из печати объём итальянского корпуса превысит 4 миллиона словоупотреблений (русская классическая литература и философия XIX–XX веков, итальянские пьесы, новеллы и повести XVIII–XIX веков).

1.6. Латышский корпус

В корпус впервые включены выровненные латышско-русские тексты, над которыми работала Н. В. Перкова. Переводы с русского на латышский пока немногочисленны — «Белая гвардия» М. А. Булгакова и рассказы Чехова; среди латышских текстов — как классическая проза (Я. Райнис, Р. Блауманис), так и литература 1970–1980-х годов, переводившаяся в советское время. Совокупный объём латышско-русского и русско-латышского компонентов — 700 тыс. словоупотреблений. Латышские тексты при их размещении были не снабжены морфологической разметкой и доступны только для текстового поиска. Отметим, что даже в Латвии задача разработки латышского морфологического анализатора до сих пор далека от удачного решения.

1.7. Немецкий корпус

В последние годы благодаря группе под руководством Д. О. Добровольского и А. А. Кретова (о которой см. раздел 1.1) в корпусе появилась русско-немецкая языковая пара, в основном состоящая из культурно значимых переводов русской классики («Капитанская дочка», «Герой нашего времени», «Война и мир» полностью, романы Достоевского и др.). Таким образом, имевшаяся в предшествующие годы асимметрия (значительный по объему англо-русский корпус, относительно небольшие русско-английский и немецко-русский корпуса при полном отсутствии русско-немецкого компонента) постепенно выравнивается. Пополнялась новыми текстами и существовавшая ранее немецко-русская языковая пара.

В корпус включены предоставленные компанией ABBYY нехудожественные тексты, переведённые с немецкого и на немецкий: производственно-технические (инструкции), учебные (санитарно-просветительские тексты), публицистика. Объем немецкого корпуса равен 7,6 млн словоформ.

Предусмотрено развитие поливариантного русско-немецкого корпуса; о существующем опыте в этом направлении см. вошедшую в сборник НКРЯ 2006–2008 статью Д. О. Добровольского. Развивается и отдельный подкорпус «Русская классика в немецких переводах» при финансовой поддержке Фонда содействия развитию интернета «Фонд поддержки интернет».

1.8. Польский корпус

Польский корпус разрабатывался в ходе совместного проекта с Варшавским университетом (руководитель с польской стороны — М. А. Лазиньский; в работе над корпусом и его программным обеспечением принимали участие П. и Н. Годлевские, Е. В. Гурина, В. И. Демидова, Е. Д. Ильина, М. Куратчик, С. О. Минлос, Г. А. Мороз, Б. В. Орехов, Е. Слободян (Выналек) и другие). Это одна из самых больших языковых пар — свыше 6,3 миллионов словоупотреблений.

Здесь хорошо представлена польская художественная проза разных периодов, от Сенкевича и Пруса до Хмелевской и Сапковского. В корпус входят также газетные статьи, законодательные тексты.

Русско-польский параллельный корпус пока не вполне репрезентативен, но в него входит, в частности, такой нестандартный для параллельных корпусов материал, как проза русского Серебряного века (Брюсов и Сологуб) в современных польских переводах.

Польский корпус размечен при помощи морфологического анализатора TaKiPi с автоматическим снятием омонимии.

1.9. Украинский корпус

Украинский корпус (работу над ним ведет в основном М. А. Шведова; в выравнивании участвовали также А. Л. Кривенко и О. А. Тищенко-Монастырская) — одна из крупнейших языковых пар в НКРЯ, вторая по объёму после английского корпуса (см. публикацию [Січінава, Тищенко-Монастирська, Шведова 2011]). Его совокупный объём составляет теперь 9,3 млн словоупотреблений. В корпусе достаточно репрезентативно собрана украинская проза, переводившаяся в разные эпохи на русский язык (от Котляревского до Жадана), представлены также выполненные в разные эпохи и с разными установками переводы с русского языка на украинский. Особо укажем на такой жанр, как украинская (и русская, переводившаяся на украинский) фантастика и приключенческая проза XX в. В украинский корпусактивно включаются нехудожественные тексты: научные, правовые, публицистические, религиозные. Контекст создания украинского корпуса несколько напоминает ситуацию с белорусским: в ситуации отсутствия большого общедоступного одноязычного корпуса украинского языка параллельный корпус в известной степени на первых порах перенимает его функции. В настоящее время вне рамок НКРЯ развиваются также украинско-польский (Н. Коцыба) и украинско-болгарский (Е. Б. Сирук, И. А. Держанский) параллельные корпуса.

1.10. Французский корпус

В формировании русско-французского компонента параллельного корпуса определяющую роль играет его поливариантный состав (подробнее см. ниже, раздел 3). Поливариантный русско-французский корпус поддержан специальным грантом (РФФИ № 12–06–33038 «Контрастивные корпусные исследования русских и фран-

цузских глагольных категорий в поливариантных параллельных текстах», руководитель Д. В. Сичинава). Основным вдохновителем и автором концепции корпуса является И. М. Зацман (ИПИ РАН), в работе над ним участвуют Н. В. Бунтман, М. Г. Кружков, Е. Ю. Лощилова, О. А. Петрушкина, Е. А. Роганова, Анна А. Зализняк и другие. На базе поливариантного корпуса построена уникальная французско-русская база данных соответствий грамматических категорий (см. подробнее ниже и публикации Loiseau et al. 2013, Бунтман и др. 2014).

В русско-французский корпус включается несколько альтернативных переводов русской классики (Гоголь, Чехов, Гончаров, Толстой) на французский язык; поливариантный состав позволяет исследовать вариативность в переводе лексики и грамматики в идентичных контекстах. Повести Гоголя «Нос» и «Шинель» в настоящее время включены в корпус в четырёх разных переводах. Наряду с поливариантными текстами в корпус включаются также и традиционные моновариантные пары (в том числе субтитры художественных фильмов). Эти моновариантные тексты выровнены Венсаном Бене (Париж, INALCO). В проекте с французской стороны участвует также Сильвен Луазо (университет Paris-XIII).

Французско-русский корпус включает моновариантные (с одним переводом) тексты, как художественные (Мопассан, Бальзак, Флобер, Госинни, Модиано), так и нехудожественные, включая предоставленные компанией ABBYY официально-деловые и технические тексты.

Совокупный объём французского компонента — 2,3 млн словоупотреблений.

1.11. Эстонский корпус

С 2015 г. в корпус входят также параллельные эстонско-русские и русско-эстонские тексты. Особая сложность для подготовки корпуса связана с плохой представленностью в Сети и других источниках электронных версий эстонской прозы (оригинальной и переводной); велась работа по сканированию и распознаванию бумажных источников. Особое влияние планируется уделить переводам «эстонского текста» русской литературы (Довлатов, Аксенов и др.) Над корпусом работали М. В. Боровикова и Е. С. Фомина (Тартуский университет).

В морфологической разметке эстонских текстов используется анализатор vabamorph, разработанный в Тартуском университете Х. Калепом и К. Муйшнек. Объем эстонского корпуса — более 400 тыс. словоупотреблений.

2. Многоязычный корпус

Популярным направлением развития многоязычных корпусов является создание так называемых «массовых параллельных текстов» [Cysouw, Wälchli 2007], включающих большое количество (не менее 10–20, а для таких массово распространяемых текстов, как Библия, и сотни) выровненных переводов некоторого текста на различные языки. Такие корпуса включают религиозные, политические, классические художественные, правовые тексты (ср. известный корпус Acquis communautaire — переводящихся на языки стран-членов ЕС правовых документов Европы) и т. п.

НКРЯ постоянно сотрудничает с построенными на основе схожей методологии (в том числе на принципе общедоступности материалов для научной работы) параллельными корпусами славянских языков, такими, как ParaSOL (ранее Regensburg Parallel Corpus, http://www-korpus.uni-r.de/ParaSol/, см. также [von Waldenfels 2006]), ASPAC (Amsterdam Slavic Parallel Aligned Corpus, авторский проект А. Барентсена, http://home.medewerker.uva.nl/a.a.barentsen/), InterCorp в рамках Чешского национального корпуса (http://www.korpus.cz/intercorp/).

В настоящее время в состав НКРЯ входит многоязычный корпус из 9 текстов («Алиса в стране чудес» и «Алиса в Зазеркалье» Л. Кэррола, «Собака Баскервилей» А. Конан Дойля, «Винни-Пух» А. А. Милна, «Код да Винчи» Д. Брауна, «Маленький принц» А. де Сент-Экзюпери, «Пиноккио» К. Коллоди, «Алхимик» П. Коэльо, «Мастер и Маргарита» М. Булгакова). Каждый из этих текстов представлен в оригиналах и в переводах (от 6 до 25 переводов, в основном на славянские языки). В некоторых случаях включено несколько переводов на один язык.

В многоязычном корпусе представлены следующие славянские языки: русский, украинский, белорусский, польский, чешский, словацкий, верхнелужицкий, словенский, хорватский, сербский, болгарский, македонский. Из неславянских языков в него входят тексты

на английском, немецком, нидерландском, шведском, итальянском, португальском, румынском, французском, греческом, латышском и литовском языках. Для части из этих языков НКРЯ поддерживает морфологический поиск (основывающийся на морфологических словарях компании «Яндекс»).

В основе многоязычного корпуса НКРЯ лежат тексты корпуса ASPAC, любезно предоставленные А. Барентсеном, и затем пополненные некоторыми текстами (например, переводами «Винни-Пуха» на балтийские языки). Кроме того, в проекте ASPAC выравнивание текстов проведено поабзацно; для нужд НКРЯ оно полуавтоматически заменено на выравнивание по предложениям.

На базе многоязычных корпусов, в частности, возможно массовое типологическое сравнение грамматических параметров языков [Dahl 2014], [von Waldenfels 2014], см. о перфекте ниже, раздел 7.

3. Поливариантный корпус

Со второй половины 2012 года команда НКРЯ совместно с исследователями Института проблем информатики РАН и французскими лингвистами (университет Paris-13, Институт восточных языков INALCO) разрабатывает поливариантный параллельный русско-французский корпус (см. подробнее [Loiseau et al. 2013, Бунтман и др. 2014]). Одновременно на его материале строится база данных поливариантных соответствий (полиэквиваленций) глагольных аспектуальных и временных форм в русском и французском языках (http://a179.ipi.ac.ru/corpora_dynasty/main.aspx).

Поливариантные параллельные корпуса — сравнительно слабо развитое направление в рамках корпусной лингвистики, и тем меньше существует собственно сопоставительных грамматических исследований, выполненных на их базе. Существует поливариантный русско-немецкий корпус на материале выполненных в разное время переводов романов Достоевского, созданный в Австрийской академии наук в Вене [Добровольский 2009], но он не находится в открытом доступе. Он используется прежде всего для исследования лексики и эволюции переводческих техник. По нескольку вариантов перевода ряда текстов на один и тот же язык (например, «Алиса в стране чудес» — шесть переводов на русский и четыре на польский) находится в ряде многоязычных параллельных корпусов, на-

пример, в Амстердамском корпусе параллельных текстов (ASPAC) и Регенсбургско-Бернском параллельном корпусе.

Для поливариантного корпуса выбираются тексты, существующие не менее, чем в двух переводах на французский язык, причём эти переводы должны быть созданы, как правило, не ранее середины XX века. Более ранние французские переводы русской литературы, особенно выполненные в XIX веке, устарели в языковом отношении, а главное, содержат много ошибок и сокращений исходного текста.

Для многих произведений русской литературы (Н. В. Гоголь, И. А. Гончаров, Л. Н. Толстой) существует несколько новых переводов, выполненных на высоком уровне. Они демонстрируют как различные подходы к художественному переводу, так и различный выбор языковых выражений, в том числе грамматических.

Пример выравнивания оригинального текста с двумя переводами («Обломов» И. А. Гончарова):

Цвет лица у Ильи Ильича не был ни румяный, ни смуглый, ни положительно бледный, а безразличный или казался таким, может быть, потому, что Обломов как-то обрюзг не по летам: от недостатка ли движения или воздуха, а может быть, того и другого.	Le teint d'Ilia Ilitch n'était ni rose, ni mat, pas véritablement pâle non plus; neutre plutôt, ou du moins ayant cette apparence, et cela peut-être parce qu'Oblomov était prématurément un peu flasque, faute d'exercice, ou faute d'air, ou faute d'exercice et faute d'air à la fois.	Le teint d'Ilia Ilitch n'était ni rose, ni hâlé, ni carrément pâle, mais indifférent ou, du moins, il le paraissait. Peut-être parce que la chair d'Oblomov était prématurément flasque: faute d'exercice ou manque d'air, peut-être l'un et l'autre.

На базе поливариантного корпуса построена контрастивная русско-французская база данных по соответствиям грамматических категорий. В базе данных эквивалентных глагольных форм объектом анализа являются «соответствия», т. е. пары функционально эквивалентных форм русского и французского языков. На материале русско-французского параллельного корпуса речь может идти о соответствиях двух типов. Это:

• «модели перевода» — множество переводов {Fn… Fn+m} для русской формы R (например, НСВ переводится как present, imparfait или в известных ограничительных контекстах passé simple);

• «стимулы перевода» — множество «стимулов» {Rn… Rn+m}, «реакцией» на которые является французская форма F (например, passé antérieur может появиться для русских слов «немедленно» или «внезапно»).

Соответствия первого типа позволяют получить новые сведения об устройстве русского языка, второго — об устройстве французского. Поскольку объектом нашего интереса является русский язык, в рамках данного проекта рассматриваются лишь соответствия в направлении «русский => французский», т. е. «модели перевода».

Второе ограничение состоит в том, что в создаваемую базу данных функционально эквивалентных лексико-грамматических глагольных форм русского и французского языков включаются только те глагольные формы русского языка, которые содержат спрягаемый глагол — verbum finitum (т. е. исключаются слова категории состояния, перифразы с глаголом «быть» и некоторые другие типы предикативных единиц).

Когда ведение базы будет отработано на тестовом материале, она будет расширена. Последующее ее расширение возможно также в направлении включения данных, полученных из параллельного французско-русского корпуса; они позволят ответить на вопрос «На какие французские «стимулы» появляется в переводе в качестве «реакции» данная русская форма?».

В рамках данного проекта рассматривается кластер глагольных категорий, который в типологической литературе получил обозначение TAM (Tense-Aspect-Mood; ср. также типологические исследования языков Европы в рамках комплексного проекта EUROTYP).

Вводится понятие «лексико-грамматической формы», которое представляет собой комбинацию значений трех категорий для русского языка (Tense, Aspect, Mood) и двух (Tense, Mood) — для французского (см. [Loiseau et al. 2001]). Понятие «лексико-грамматической формы» (ЛГФ) близко по своему содержанию к понятию «конструкции» — в том значении, которое придается этому термину в Грамматике конструкций [Goldberg 1995, 2006], т. е. это может быть как набор значений грамматических и семантических категорий, задающий класс удовлетворяющих данным требованиям реальных форм (т. е. «грамматическая конструкция» в традиционном понимании), так и единичная лексема или словоформа. Понятие лексико-грамматической формы представляется

наиболее адекватным инструментом контрастивного анализа поливариантного параллельного корпуса.

4. Расширение жанрового многообразия корпуса

В параллельный корпус в настоящее время активно включаются нехудожественные тексты. Преимущество их использования — в меньшей степени переводческой вольности на всех уровнях передачи текста по сравнению с художественным переводом, в лучшей представленности в текстах современной лексики, терминологии. К сложностям следует отнести неопределенность в ряде случаев языка оригинала и языка перевода (например, в двуязычных инструкциях к лекарствам, текстах Интернет-СМИ и т. п. о направлении перевода можно строить только более или менее правдоподобные предположения), а также сомнительную лингвистическую корректность использования текстов, возникших в результате вручную отредактированного машинного перевода (это не редкость, например, в двуязычных украинских СМИ; однозначной диагностике такие случаи не поддаются). Существенное подспорье Корпусу оказал проект ABBYY Lingvo, с любезного согласия которого в английский, немецкий и французский корпуса были включены нехудожественные тексты разных жанров.

Сейчас в параллельных корпусах 3,2 млн словоупотреблений нехудожественных текстов (4,6% совокупного объёма). Уже представлены те же основные сферы функционирования текстов, что и в основном одноязычном русском корпусе: официально-деловая, обиходно-бытовая, производственно-техническая, публицистика, религиозная, учебно-научная. В различных языковых парах они представлены неравномерно; в «малых» парах (армянском, испанском, латышском, эстонском корпусах) нехудожественных текстов пока нет. Больше всего удельный вес нехудожественных текстов в белорусском (9%) и в украинском (11%) корпусах.

Здесь же можно упомянуть об использовании «неканоничной» художественной литературы: литературно обработанного фольклора (народных сказок), массовых жанров вроде детектива или фантастики. Художественные тексты таких жанров пополняют также прежде всего украинский и белорусский корпуса.

5. Средства разметки

5.1. Выравнивание

Для параллельных корпусов НКРЯ были разработаны специальные программные средства выравнивания и разметки. Команда, работающая над английским и немецким корпусом, использует программу «ПарТекс» (идея А. А. Кретова, руководство И. Е. Ворониной, программирование Д. Спесивцева), имеющую на входе два параллельных текста (оригинал и перевод). Выравнивание осуществляется на уровне предложений. Программа выдает на выходе синтезированный текст, в котором последовательно за каждым предложением оригинала следует соответствующее предложение перевода. В таком синтезированном тексте поиск интересующих пользователя слов может осуществляться штатными средствами обычных текстовых редакторов. В программе «ПарТекс» для поиска может быть «подстрока в строке» и на выход подается текстовый файл, в который входят все пары предложений оригинала и перевода, содержащие заданную последовательность символов. Возможен поиск как английских или немецких, так и русских слов или словосочетаний.

Одна из наиболее существенных трудностей выравнивания заключается в том, что авторское членение текста на предложения и абзацы не всегда выдерживается в тексте перевода. Кроме того, в разных языках (а иногда и разных изданиях) приняты различные способы графического оформления, что иногда затрудняет определение границ предложения в автоматическом режиме. Ср., например, различные способы оформления переходов от прямой речи персонажей к авторским ремаркам. В таких случаях результаты автоматического выравнивания нуждаются в коррекции, осуществляемой вручную. Программа позволяет обнаружить в параллельных текстах асимметрию такого рода. Результаты работы программы ПарТекс далее просматриваются и дополнительно корректируются вручную (Д. О. Добровольским и Д. В. Сичинавой, в последние годы также М. А. Хохловой из СПбГУ / ИЛИ РАН).

Для выравнивания остальных параллельных корпусов (кроме английского и немецкого) используется программа «Евклид» (идея Д. В. Сичинавы, программист Т. А. Архангельский, учитывались за-

мечания и предложения разметчиков из разных стран, прежде всего В. А. Кощенко [Беларусь] и М. А. Шведовой [Украина]). Эта программа является графическим интерфейсом пользователя (GUI) к доступной в открытом режиме программе выравнивания текстов HunAlign [Varga et al. 2005].

Входящий стандарт программы — файл в текстовом формате (ANSI или Юникод) с любым числом абзацев, разбиений строки, пробелов в начале строк и т. д. По запросу пользователя загружается несколько текстов — оригинал и перевод(ы), после чего они проходят автоматическую предобработку: удаляются пустые строки, лишние пробелы, каждое предложение переносится на новую строку, при этом учитываются основные сочетания знаков препинания и типы границы предложения.

Затем запускается программа HunAlign, при помощи статистического механизма (информация Гейла-Чёрча о длине предложений и автоматическое составление словаря на её базе) автоматически выравнивающая тексты попарно полностью, и выход её загружается в графический интерфейс программы «Евклид» для ручного постредактирования.

Программа HunAlign, основываясь на ряде статистических весов (длина предложения, совпадение символов, структура знаков препинания и т. п.) приписывает каждой выровненной ею паре предложений определённый коэффициент вероятности выравнивания. Если он ниже нуля, значит, выравнивание данных отрезков текста маловероятно. Кроме того, она автоматически склеивает последовательности предложений на одном языке, соответствующие, с его точки зрения, друг другу. Места склейки в выходном формате при этом указываются.

Пары предложений с отрицательным коэффициентом выравнивания и все склеенные последовательности предложений считаются сомнительными отрезками, нуждающимися в первоочередной проверке вручную. Для графической чёткости коэффициенты в сомнительных отрезках выделяются в программе цветом. При помощи специальной опции в программе «Евклид» скрываются все предложения, кроме сомнительных отрезков. Они просматриваются редактором и в случае ошибки выравнивания (а также в случае сохранения в файле лишней информации — например, записанных в текстовом виде выходных данных и т. п.) редактируются вручную

при помощи быстрых механизмов программы. Предусмотрены следующие автоматизированные операции (в виде графических кнопок при каждой строке, некоторые из которых открывают отдельное диалоговое окно):

добавление пустой строки;

удаление пустой строки;

перенос предложения в соседнюю строку;

перенос части склеенного предложения — т. н. «обрезка» — в соседнюю строку.

При необходимости ближайший контекст сомнительных отрезков может быть раскрыт (двойным щелчком по скрытой строке) без раскрывания всего остального текста. Затем для просмотра и редактирования можно сделать доступным весь остальной текст (без сомнительных отрезков) и отредактировать его по тому же принципу.

5.2. Формат

Выровненный текст можно сохранить в формате XML (кодировка Юникод), как в окончательном корпусном, так и в промежуточном формате, доступном для дальнейшего редактирования. При сохранении выровненного текста следует указать язык оригинала и перевода. Они будут сохранены в XML в составе соответствующих тегов. В сохранённом XML пары предложений пронумерованы (им сопоставлены уникальные номера) для облегчения ссылок.

Программа «Евклид» позволяет также добавить к каждому сохранённому тексту метатекстовую информацию (метаразметку), которая записывается в качестве отдельной строки в электронную таблицу (документ, доступный для редактирования программами типа Excel). Пользователь заполняет в специальной форме название текста и имя автора в оригинале и переводе, дату создания текста; язык оригинала и перевода сохраняются автоматически исходя из ранее сохранённой информации.

Выровненные пары предложений объединяются при помощи XML-тега <para>. Тег имеет атрибут ID, в который записывается номер предложения.

Предложения объединяются при помощи XML-тега <se>. Тег имеет обязательный атрибут lang, где указан язык текста, например, <se lang=ru> для русского или <se lang=be> для белорусского.

5.3. Метаинформация

Метаинформация сохраняется в отдельном файле (электронной таблице) в формате CSV (значения ячеек разделены при помощи знака «точка с запятой»). Таблица заполняется при помощи соответствующей формы программы «Евклид». Таблица включает в себя следующие поля метаинформации:

1) название текста в оригинале;
2) год создания текста;
3) имя автора в оригинале;
4) год рождения автора;
5) название текста в переводе;
6) имя автора в переводе;
7) имя переводчика (на языке перевода);
8) дата перевода;
9) сфера функционирования текста (художественная, публицистика...);
10) язык оригинала;
11) язык перевода.

5.4. Маркирование вольности перевода

Важно отметить, что художественный перевод советского времени с русского языка на языки народов СССР и в обратном направлении (а именно переводы этой эпохи лежат в основе нынешнего состава украинско-русского и белорусско-русского параллельных корпусов) отмечен высокой степенью вольности на разных уровнях, и при составлении и анализе параллельного корпуса это постоянно приходится учитывать. Причины этого разнообразны — общая установка ряда переводчиков, роль автора, участвующего в творческом пересоздании авторизованного перевода (в частности, он может и сам выступать переводчиком), а нередко и цензурные причины. Большое значение имеет корпус для исследований по теории и практике перевода, например, по передаче заглавий, идиом, различного рода идеологем (так, анализ переводов советского времени с украинского на русский язык показывает, что в них систематически различными переводчиками устранялась религиозная фразеология и оскорбительная в современном языке [но далеко не всегда для

авторов XIX в.] этнонимия — вместо неба крестьяне апеллировали к справедливости, а *жид* и *москаль* достаточно регулярно представали, к примеру, *ростовщиком* и *солдатом*). Вольности и ошибки различного уровня фиксируются и в переводах с участием других языков, хотя и не в таких масштабах, но также систематически.

Для указания наличия той или иной вольности в переводе,— пропусков, добавлений, замены текста с изменением смысла,— в корпусе используется специальный уровень разметки — атрибут **loose**. Значения атрибута следующие:

Добавление: <se lang=rus loose=add>
Пропуск: <se lang=rus loose=omit>
Изменение: <se lang=rus loose=change>

Пример из украинского рассказа «Невольница» Марко Вовчок (автоперевод); выделены добавленные в переводе фрагменты:

<para id="84">
<se lang="uk">Ще гаразд сонечко не підбилося у небі, вже скрізь взброювалися як спроможність, коней виводили, сідлали,— в кого чого не було — біг той питаючи, шукаючи; козачки допомагали й кохали як мога; діти клопоталися, начеб їм на турка перед вести йшлося…</se>
<se lang="ru" loose="add">Еще солнышко хорошенько не взошло на небе, уж повсюду в городе вооружались чем кому бог послал, всюду коней выводили, седлали — кому чего недоставало, тот бежал, спрашивая, ища, отыскивая; козачки помогали, как только в силах были; *надо сказать, что* они и любили тогда, как только в силах были: *не раз вы, верно, деточки, замечали сами, что вас с большей нежностью гладили по головке, когда вы, по детской неосмотрительности, падали, например, в грязь — каково ж должно быть чувство к тому, кто идет честно на смертный бой?* Дети хлопотали и *суетились*, словно им приходилось самим в передовом отряде на турок идти…</se>
</para>

Пример из «Смерти Ивана Ильича» Л. Н. Толстого во французском переводе М. Оффмана; выделен опущенный в переводе фрагмент.

<para id="353">
<se lang="ru">*Было лучшее общество, и* Иван Ильич танцевал с княгинею Труфоновой, сестрою той, которая известна учреждением общества «Унеси ты мое горе».</se>
<se lang="fr" loose="omit">Ivan Illitch dansa avec la princesse Troufonova, la propre sœur de l'illustre fondatrice de la société « Foin de peines! ».</se>
</para>

Пример из «Похищения чародея» К. Булычева в белорусском переводе Р. Саматыи; выделены изменённые и добавленные фрагменты:

<para id="90">
<se lang="ru" variant_id="0" loose="change">*Речь шла* об опустении рек и лесов, о том, что некий купец еще до революции возил с холма камень в Полоцк, чем обкрадывал культурное наследие, о том, что население этих мест смешанное, потому что сюда все кому не лень ходили, о том, что каждой деревне нужен музей…</se>
<se lang="be" variant_id="1" loose="add, change">*Яго доўгі маналог* быў пра горкі лёс рэк і лясоў; пра тое, што якісьці купец яшчэ да рэвалюцыі вазіў з пагорка камень у Полацак, абкрадаючы *гэтым самым* культурную спадчыну; пра тое, што насельніцтва гэтых мясцін мяшанае, бо сюды прыходзілі *ваяваць усе, хто хацеў*; што кожнай вёсцы патрэбны музей…</se>
</para>

В рамках поливариантного корпуса разработана более тонкая разметка неточных соответствий перевода (см. подробнее Сичинава 2013, Loiseau et al. 2013).

6. Пример использования параллельного корпуса для количественного изучения лингвоспецифичной лексики[1]

Использование параллельных корпусов для сравнительных исследований — предмет уже обширной исследовательской литера-

[1] Данное исследование поддерживалось грантом РФФИ № 13-06-00403 «Контрастивное корпусное исследование специфических черт семантической системы русского языка» (руководитель Анна А. Зализняк).

туры, назовём, прежде всего, специальный номер журнала STUF под редакцией М. Сисоу и Б. Вельхли [Cysouw, Wälchli (eds.) 2007]. При переводе лексики, как и грамматических конструкций (чёткую границу между ними не всегда можно провести), существенную роль играют **модели перевода** и **стимулы перевода**. Обычно для конструкции А в языке В имеется (реально представлено в текстах) несколько неслучайно повторяющихся соответствий — моделей перевода (translation patterns; ср. сборник [Hasselgård, Oksefjell (eds.) 1999]. Практически можно утверждать, что, имея дело с переводными текстами и располагая сколько-либо большим корпусом, мы наверняка можем выделить несколько моделей перевода конструкции языка А в языке В. Стимулом перевода назовем конструкции в языке оригинальных текстов, соответствующие определённой конструкции в языке перевода («что стимулирует появление конструкции X в переводе на язык В?»).

Параллельный корпус может использоваться для точного определения и выявления так называемой лингвоспецифичной лексики (единиц, как нередко утверждается, т. н. «языковой картины мира»), которая «трудно» или особенно неоднозначно переводится на другие языки (ср. [Добровольский 2009]). Вокруг этого и схожих понятий, как известно, ведётся оживлённая дискуссия (ср. резко критикующие любые исследования «языковой картины мира» публикации, собранные в книге [Павлова (сост.) 2013], а также взвешенный обзор дискуссии в [Руссо 2014]). Представляется, что некоторая попытка точных формализуемых определений здесь была бы как минимум не лишней.

Возможных мер «разброса различных переводов» некоторой лексемы можно предложить несколько, в частности:

отношение абсолютной частоты самой частотной модели перевода (F (Mmax)) к количеству различных эквивалентов (NumM);

средняя частота вхождений на один эквивалент (F (O)/NumM, где F (O) — частотность данного слова в оригинале);

отношение абсолютной частоты самой частотной модели перевода к частоте второй (F (Mmax)/F (Msec));

отношение абсолютной частоты самой частотной модели перевода к общему количеству вхождений (F (Mmax)/F (O)).

Наконец, существуют несколько самых общих статистических мер разброса (diversity indexes), из которых мы используем две:

энтропию дискретного распределения относительной частотности каждой модели перевода ($-\sum (F (M_i)/F (O))\log2 (F (M_i)/F (O))$ и более простой индекс Герфиндаля (имеющий также ряд других названий), первоначально использовавшийся для оценки уровня монополизма в экономике ($\sum (F (M_i)/F (O))^2$; минимальное значение: 1/NumM, максимальное 1[1]). Высокое значение энтропии соответствует низкому индексу Герфиндаля и большому разбросу значений.

Для «лингвоспецифичного» слова (если таковые существуют) предполагается, что моделей перевода будет много, в среднем на каждую будет приходиться сравнительно немного контекстов, а частота самой частотной из них не будет сильно отличаться от остальных (и он будет занимать лишь небольшой процент от общего числа соответствий). Энтропия будет достаточно велика, а индекс Герфиндаля — низок.

Данные показатели имеет смысл вычислять также для стимулов перевода (в этом случае предполагаемая лингвоспецифичная лексика появляется при переводе; соответственно вместо буквы M выше выступает S, а вместо частотности в оригинале F (O) — частотность в переводе F (T))

Кроме того, имеет значение отношение частоты лексемы (на миллион слов) в оригинальных текстах на языке L (F (O)ipm) к частоте ее же (на миллион слов) в переводных текстах на этом же языке L (F (T)ipm). Ожидается, что лингвоспецифичная лексика скорее появится при создании оригинального текста, чем при переводе.

Данные показатели можно получить и сравнить для нескольких языков (например, для лексики русского языка использовать русско-английский, русско-украинский и другие корпуса).

Проверим эти предположения на материале нескольких лексем, отнесенным в [Зализняк, Левонтина, Шмелев 2012] к лингвоспецифичным, и нескольких лексем, которые такими обычно не считаются.

Однокоренные модели перевода обычно засчитываются как одна модель (bore, boredom, boring, bored); то же относится к случаям, когда в некоторых вхождениях имеются также другие слова, например, recklessness приравнивается к reckless jockeying. Повторяющиеся в одном тексте идиоматизмы (фразеологизмы типа *заливать то-*

[1] Существует также нормализованный индекс Герфиндаля, независимо от объёма множества принимающий значения от 0 до 1: (∑(F(Mi) /F(O)) — 1/NumM)/(1-1/NumM).

ску, прозвища типа *Смертная Тоска*), внутренние цитаты и т. п. явно зависимые вхождения считаются только один раз. Случаи пропуска слова в переводе или, напротив, привнесения переводчиком («отсебятины»: *Fredrica's cat watched her from the high window* / *Из окна под самой крышей дома за ней* ***с тоской в глазах*** *наблюдал кот Фредерики*; *He was propped behind the wheel of his shiny yellow Lincoln, talking about his grandmother* / *Он сидел за рулем своего сверкающего желтого «Линкольна» и,* ***уставившись куда-то в пространство остекленевшими глазами****, о чем-то разговаривал со* [sic! ср. about] *своей бабушкой*) не учитываются. Иногда неясно, какое слово соответствует какому или же перевод «распределен» по нескольким; например, во фразе *И как чудна она сама, эта дорога: ясный день, осенние листья, холодный воздух… покрепче в дорожную шинель, шапку на уши, тесней и уютней прижмемся к углу!* [Н. В. Гоголь. Мертвые души (1835–1852)] // *and how interesting for its own sake is a highway, should the day be a fine one (though chilly) in mellowing autumn, press closer your travelling cloak, and draw down your cap over your ears, and snuggle cosily, comfortably into a corner of the britchka.* [Nikolay Gogol. Dead Souls]. Все три корня *snug*, *cosy* и *comfort* выступают в других текстах для передачи слова *уют(ный)*.

Если сравнить лексемы *удаль*, *тоска* и *страсть* (последнее в значении «желание, стремление, чувство», а не омоним со значением «страх»), то оказывается, что у первых двух очень сильный разброс лексических соответствий в параллельных английских текстах. В оригинальных текстах слово *удаль* встретилось 4 раза и переведено каждый раз по-новому, соответственно, значения F (Mmax)/F (Msec) = F (O)/NumM = 1 (минимально возможное значение); F (Mmax)/NumM = F (Mmax)/F (O) = 0,25. В переводных текстах слово удаль встретилось 11 раз, 2 стимула встретились дважды, 9 по одному. Соответственно, это даёт F (Smax)/NumS = 0,22, F (Smax)/F (T) = 18%, F (Smax/Ssec) = 1, F (T)/NumS =1,2. В оригинальных текстах (на миллион слов) слово удаль встречается в 1,6 раз чаще, чем в переводных. Показатель энтропии H для моделей перевода — 2, для стимулов перевода — 3,4. Индекс Герфиндаля для моделей перевода — 0,25 (минимально возможный для 4 значений; нормализованный тем самым равен 0), для стимулов перевода — 0,14 (нормализованный 0,05).

Более частотная лексема *тоска* переводится на английский

22 различными способами и соответствует 66 различным английским корням в оригинале (различие связано, в частности, с тем, что англо-русский корпус в 4,6 раза больше русско-английского). Этот огромный список включает, в частности, новые заимствования: *ennui* из французского и *angst* из немецкого, что может уже косвенно указывать на специфику русского слова. При переводе на английский, благодаря индивидуальным предпочтениям переводчиков, выделяются два слова: *anguish* (его предпочитают переводчики Булгакова Пивир и Волохонская) и *misery* (его предпочитает переводчица Чехова К. Гарнетт; в частности, существует одноименный рассказ Чехова, где заглавное слово *тоска*/misery встречается семь раз). При том, что средняя частота соответствия, естественно, выше, чем у редкого слова *удаль* (3,2), выше и отношение частоты лидирующего варианта к количеству остальных (0,72), остальные показатели близки: F (Mmax)/F (Msec) = 1, F (Mmax)/F (O) = 23%. Для стимулов перевода, где лидерство — за основой long-, а на втором месте — предлагаемое А. Вежбицкой в качестве основного эквивалента yearning и его производные, разрыв между двумя местами больше, чем в русско-английском: F (Smax/Ssec) = 1,92, очень близкая средняя частота соответствия (3,02) и гораздо более низкий рейтинг лидирующего перевода — 13%. Показатели энтропии вдвое ниже, чем у *удаль* (для моделей перевода — 1,08, для стимулов — 1,6), зато ниже и индекс Герфиндаля: 0,13 и 0,04 (нормализованный 0,12 и 0,03) соответственно. Слово *тоска*, как и слово *удаль*, встречается в оригинальных текстах в 1,6 раз чаще, чем в переводных.

Совсем иная картина открывается при анализе профиля слова *страсть*, примерно столь же частотного в русском тексте (26 на миллион), как *тоска* (24 на миллион). В 85% случаев это слово переводится с русского как passion, разброс стимулов больше — на самый частотный стимул (разумеется, тоже passion) приходится 34% случаев. Дело в том, что слово *страсть* несколько чаще встречается в переводных текстах — в 1,3 раза; оно обслуживает, в частности, такую английскую лексику, как lust (‘~похоть’), obsession (‘~одержимость’) и другие. Отметим, что общие статистические показатели демонстрируют резкое различие между оригинальными и переводными текстами: показатель энтропии 0,31 для моделей, но достаточно высокий (0,99) для стимулов; индекс Герфиндаля 0,73 для моделей и 0,16 (нормализованный 0,14) для стимулов. Та-

ким образом, слово *страсть* ведёт себя (судя по моделям перевода в русском оригинале) не похоже на лингвоспецифичное слово, но (как показывают стимулы русского перевода) вместе с тем является общим термином (umbrella term) для целого ряда английских наименований чувств.

Интересный материал представляют родственные квазисинонимы *простор* и *пространство*, первый из которых нередко считается знаком русской «языковой картины мира», а второй — как правило, нет. Слово *простор* неоднократно привносится переводчиком как часть речения (сложного предлога?) *на просторах, в просторе*: in the ocean — *на просторах* океана; такие случаи, в соответствии с общей нашей методикой, при подсчётах не учитывались, но отметить их заметное количество необходимо. Обнаруживается значительный разброс моделей перевода слова *простор* (самый частотный перевод, space и его производные — 21% вхождений, средняя частота переводов — 1,5 слова, энтропия 0,92, нормализованный индекс Герфиндаля 0,07); еще сильнее это выражено для стимулов — самый частотный перевод 19%, средняя частота 2,6, энтропия 1,34, нормализованный индекс Герфиндаля 0,06. Слово *пространство* имеет существенно более низкий разброс соответствий: самый частотный перевод (тоже space) — 55% среди моделей и 66% среди стимулов; частота среднего перевода — 3,8 среди моделей и 4,6 среди стимулов; энтропия 0,72 среди моделей и 0,63 среди стимулов, нормализованный индекс Герфиндаля соответственно 0,32 и 0.44. При этом оба слова чаще встретились в переводных текстах, чем в оригинальных: *пространство* — примерно в два раза, *простор*, что отчасти предсказуемо, лишь в 1,3 раза.

Рассмотрим также еще два слова, часто считающиеся лингвоспецифичными: *уютный/уют* и *пошлый/пошлость*. Показатели слова *уют* напоминают показатели слова *страсть*: переводчики с русского используют резко ограниченный круг соответствий (cozy/cosy, comfort и snug — первое из них характерно для переводчицы Стругацких А. Буис и уже упомянутых Пивира и Волохонской, последнее — для К. Гарнет), в то время как переводчики с английского привносят слово *уют* как общий umbrella term для положительных характеристик «атмосферы» такого рода, в результате чего оно почти вдвое чаще встречается в переводных текстах. Лингвоспецифичность тут если и есть, то особого рода, и зависит от направления

перевода. Действительно, на невыразительные показатели для моделей перевода (37% для самой частотной модели, средняя частотность модели 7,5, энтропия 0,52, нормализованный индекс Герфиндаля 0,29), приходится заметно большая энтропия при переводе на русский (0,92, нормализованный индекс Герфиндаля 0,20), при том, что лидирующий стимул перевода даёт даже более высокий показатель, чем лидирующая модель (44%).

Слово *пошлый (пошлость)* обнаруживает ряд признаков этого же типа (ограниченный круг выбора в переводах с русского и более широкий разброс стимулов при переводах с английского), но всё же оно заметно ближе к «хорошим» лингвоспецифичным словам типа *удаль* и *простор*. Среди моделей перевода обнаруживается достаточно высокая энтропия (0,8), но нормализованный индекс Герфиндаля даёт 0,26 (почти в половине случаев используется слово vulgar). По стимулам перевода разброс, как обычно, выше: энтропия вырастает до 1,1, нормализованный индекс Герфиндаля падает до 0,06, а на vulgar приходится лишь пятая часть соответствий. Наконец, *пошлость*/*пошлый* существенно чаще встречается в оригинальных текстах (почти в пять раз), и это наиболее яркая его лингвоспецифичная черта.

Любопытно привлечь также материал близкородственного языка — украинского и сравнить меру специфичности украинских соответствий с английскими соответствиями; соответствующие показатели для 9-миллионного параллельного русско-украинского корпуса также вполне надёжны. Оказывается, что только два слова из этого перечня могут, исходя из статистических критериев, считаться лингвоспецифичными для русского по сравнению с украинским — *пошлый (пошлость)* и *удалой (удаль)*: как для моделей, так и для стимулов перевода обоих этих слов энтропия выше 0,65, а средний нормализованный индекс Герфиндаля — 0,18. Слово *тоска* даёт средние показатели разброса, несмотря на большое количество разных соответствий, а слова *уют(ный)* и *страсть* уже имеют четко выраженные ведущие переводные эквиваленты (*затишок/затишний* и *пристрасть* соответственно). Показательно, что слово *простор*, представавшее как «хорошее» лингвоспецифичное при сопоставлении с английским, при сопоставлении с украинским даёт, наоборот, самый низкий разброс соответствий: дело в том, что в украинском языке имеется его точное этимологическое соответ-

ствие *простір*, семантически отвечающее не только русскому *простор*, но и *пространство*. По моделям перевода слово *пространство* при этом имеет несколько больший разброс, чем *простор* (и даже *уют* и *страсть*), но по стимулам перевода выглядит примерно так же, как и *простор*.

Итак, применение статистических методов к выделению лингвоспецифичной лексики на параллельном корпусе кажется перспективным и выявляет, кроме того, ряд общих свойств переводных текстов. Например, разброс моделей перевода как общее правило ниже, чем стимулов. Выявлены слова, типа *уют* или *страсть*, для которых эта разница особенно сильна и которые используются именно при переводе как соответствие большому ряду английских эквивалентов. Применение этой методики к целому ряду романских, германских и славянских языков позволит выделить кластеры специфичной для тех или иных языков лексики.

7. Пример использования параллельного корпуса для грамматических исследований: перфект в параллельном корпусе

Параллельный корпус является хорошим материалом для исследования такой типологически нетривиальной и варьирующей даже в близкородственных языках категории, как перфект[1]. Об исследовании перфекта в «массовом» параллельном корпусе библейских переводов ср. также выполненную независимо от нашей работу [Dahl 2014]; подробнее материалы нашего исследования изложены в работе [Сичинава 2016].

На материале анализа переводов «Винни-Пуха» А. А. Милна на другие германские языки видно, что английский перфект почти всегда переводится при помощи германского аналитического перфекта (с точностью до выбора вспомогательного глагола в немецком и нидерландском), но при этом имеется и ряд отклонений, прежде всего, связанных с небазовыми значениями перфекта. (О семантике перфекта в типологическом освещении, и особенно в европейских языках, см., прежде всего, [McCoard 1978], [Dahl, Hedin 2000], [Lindstedt 2000]).

[1] Работа поддерживалась грантом № 10-04-00168 «Механизмы изучения перфекта в типологическом освещении», руководитель проекта Ф. А. Елоева.

Перфект пассива с результативным значением может переводиться презенсом (в данном примере — на немецкий):

EN: The flood-level has reached an unprecedented height [WP: IX]
DE: Der Pegelstand issst (PRAES) unverhältnisssmäßßig hoch.
'Уровень паводка достиг небывалой высоты'

Перфект со значением оценки ситуации («оказался») переводится на немецкий простым претеритом, на нидерландский — конструкцией без глагола («так глупо меня (можно) ввести в заблуждение»). Из германских языков только шведский, язык с «сильным перфектом», сохраняет форму. Романские языки и македонский сохраняют перфект, но в балтийских и болгарском выступают другие времена:

EN: «I have been Foolish and Deluded,» said he [WP: III]
DE: «Ich war (PRAET) ein verblendeter Narr», sagte er.
NL: «Stommerd om me zo op een dwaalspoor te laten brengen,» zei hij.
SV: — Jag har varit enfaldig och blivit lurad, sa han.
ES: «He sido Crédulo y Estúpido», dijo.
IT: «Sono stato Stupido e Ingenuo» disse.
LT: — Buvau (PRAET) žioplas Apsigavėlis,— pridūrė jis.
LV: — Es biju (PRAET) muļķis un piekrāpos,— viņš bēdājās.
MK: «Колку сум бил глупав и забудален» рече.
BG: — Аз съм (PRAES) Глупав и Измамен…
'— Я был Глуп и Сбит с толку,— сказал он'

Претеритом в немецком передан и перфект «расширенного настоящего», охватывающего точку отсчёта. Наиболее ярко это выражено в примере, где говорящий противопоставляет настоящее и перфект «расширенного настоящего» при помощи наречия *so far* 'до сих пор' (перевод презенсом, таким образом, невозможен). В немецком и нидерландском выбран претерит, при этом в шведском перфект сохранён:

EN: «Friends», he said, «including oddments, it is a great pleasure, or perhaps I had better say it has been a pleasure so far, to see you at my party». [WP: X]
NL: «Vrienden», zei hij, «vrienden en alles wat daarbij hoort, het is mij

een groot genoegen, of misschien moet ik zeggen, tot nu toe was (PRAET) het een groot genoegen om jullie allen op mijn partijtje te zien».

DE: «Freunde», sagt er, «und sonstiges herumwuselndes Kroppzeug eingeschlossen, es ist ein großes Vergnügen oder vielleicht sollte ich eher sagen, Es war (PRAET) bisher ein großes Vergnügen euch auf meiner Party zu sehen».

'Друзья,— начал он,— друзья мои… включая прочих! Для меня большая радость — во всяком случае, до настоящей минуты было большой радостью — видеть вас на моём Пиргорое'.

Отметим, что большинство отмеченных отклонений связано с бытийным глаголом, перфект которого в немецком (*ich bin gewesen*) вообще существенно более редок, чем претерит (*ich war*).

При обращении к англо-русскому параллельному корпусу оказывается, что в большинстве случаев английский перфект передаётся русским претеритом без каких-либо дополнительных лексических средств. Перфекту с обстоятельством ограниченного периода времени в русском соответствует глагол СВ длительно-ограничительного способа действия (пердуратив) с приставкой *про-*:

«Vicky», he said carefully. «I have driven fifteen hundred miles on turnpikes since we left Boston».

'— Вики, я проехал (// ?ехал) по шоссе пятнадцать тысяч миль <…>'. [Стивен Кинг. Дети кукурузы / Пер. не указан].

Обращает на себя внимание использование несовершенного вида прошедшего времени (выражающего одновременность ситуации) для передачи значения «результатив-в-прошедшем» (в составе инфинитива); СВ *осталось* в этом контексте менее вероятен. При этом фраза перестроена: вместо 'что я оставила' — *что у меня оставалось*. Если бы агентивная конструкция оригинала была сохранена (*что я оставила достаточно времени*), то здесь как раз невозможен был бы НСВ (**что я оставляла*…).

(26) I strolled for a bit, happy to have left enough time to get as lost as I was, and finally ducked into a deli for a cup of coffee. [Lauren Weisberger. The Devil Wears Prada (2003)]

'Я немного поблуждала, радуясь, что у меня еще оставалось на это

время, и наконец толкнулась в закусочную, чтобы выпить чашку кофе'. [Лорен Вайсбергер. Дьявол носит Прада / М. Маяков, Т. Шабаева].

Несовершенный вид выступает в случае, когда перфект сочетается с итеративным значением (фактически эта функция близка к таксисной):

Sometimes I try to call up old girl friends on the telephone late at night, after my wife has gone to bed. [Kurt Vonnegut. Slaughterhouse-Five Or The Children's Crusade (1969)]

'Иногда поздно ночью, когда жена уходит спать, я пытаюсь позвонить по телефону старым своим приятельницам' [Курт Воннегут. Бойня номер пять, или Крестовый поход детей / Р. Райт-Ковалёва]

Перфекту со значением «расширенного настоящего» (актуальной ситуации, охватывающей момент речи) в русском отвечает презенс:

And now for four or five lifetimes of men the Godkings have ruled all the four lands together, and made them an empire. [Ursula Le Guin. The Tombs of Atuan (1971)]

'И вот уже четыре или пять поколений Божественный Король правит всеми Четырьмя Странами, объединенными в Империю' [Урсула Ле Гуин. Гробницы Атуана / Пер. не указан]

Перфект прогрессива передаётся несовершенным видом (настоящего либо прошедшего времени) в актуально-длительном значении. Отмечено такое лексическое средство для передачи охватывающего точку отсчёта перфекта прогрессива, как глаголы типа *начать, решить*, указывающие на актуальность для текущего момента начального предела ситуации:

Perhaps people have been celebrating Bonfire Night early — it's not until next week, folks! [Joanne Kathleen Rowling. Harry Potter and the Sorcerer's Stone (1997)]

'Кажется, народ уже **начал** праздновать день Порохового Заговора — рановато, господа, он будет только на следующей неделе!' [Дж. К. Роулинг. Гарри Поттер и Волшебный камень / М. Спивак].

Well, it's a new thing the boss has been trying. [Stephen King. The Lawnmower Man (1975)]

— Это новая методика, которую **решил** испробовать наш босс. [Стивен Кинг. Газонокосильщик / Пер. не указан]

Перфект пассива в результативном значении передаётся стандартно выражающим «результатив-в-настоящем» русским презенсом пассива (ср. Князев 1983):

The golden rule of the crime scene, people, is don't touch anything until it has been studied, photographed and charted. [Michael Connelly. City Of Bones (2002)]

'Ребята, золотое правило осмотра места преступления заключается в том, что нельзя притрагиваться к находке, пока она не осмотрена, не сфотографирована и не отмечена на плане' [Майкл Коннели. Город костей / Д. Вознякевич].

Ср. не вполне обычно звучащий перевод системного сообщения: *Установка Windows была успешно завершена* в соответствии с *has been (*was) successfully completed*. Действительно, по-русски претерит пассива употребляется скорее в антирезультативных контекстах или по крайней мере в нарративе, чем в контексте «свежих новостей» (hot news) и достигнутого в настоящем результата (ср., впрочем, поэтическое: *Москва, спалённая пожаром, французу отдана*).

Отметим употребление презенса для передачи перфекта при глаголе (пересказываемой) речи. Такая сочетаемость в контекстах пересказа текста характерна для русского языка (*говорят, что…, ты пишешь, что*…, сюда же, вероятно, «вневременной презенс» типа *Аристотель делит формы государства на правильные и неправильные*). При этом перфект прогрессива (описывающий предшествующие моменту сообщения события) передан русским претеритом НСВ:

And finally, bird-watchers everywhere have reported that the nation's owls have been behaving very unusually today. [Joanne Kathleen Rowling. Harry Potter and the Sorcerer's Stone (1997)]

'И в завершение нашего выпуска, орнитологи страны сообщают, что сегодня повсеместно наблюдалось крайне странное пове-

дение сов'. [Дж. К. Роулинг. Гарри Поттер и Волшебный камень / М. Спивак]

Ср. аналогичную корреляцию презенса и перфекта в переводе с русского на английский:

Ведь целый ряд очень видных ученых полагает, что находки в зонах посещения способны изменить весь ход нашей истории. [А. Н. Стругацкий, Б. Н. Стругацкий. Пикник на обочине (1971)]

'After all, many very important scientists have proposed that the discoveries made in the Visitation Zones are capable of changing the entire course of our history' [Arkady Strugatsky, Boris Strugatsky. Roadside Picnic / Antonina W. Bouis]

Перфекту в экспериенциальном значении соответствует несовершенный вид в общефактическом значении:

Until recently antimatter has been created only in very small amounts (a few atoms at a time). [Dan Brown. Angels and Demons (2000)]

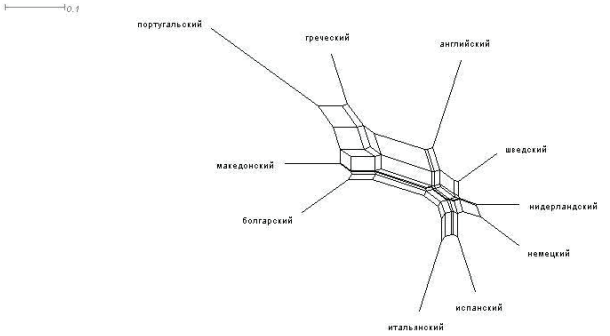
'До недавнего времени антивещество получали лишь в мизерных количествах (несколько атомов за один раз)' [Дэн Браун. Ангелы и демоны / Г. Косов]

"God damn it; who the hell has been tampering with this?" [Isaac Asimov. The Gods Themselves (1972)]

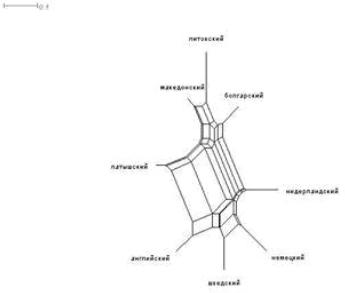
'— Черт подери! Какой сукин сын трогал эту колбу?' [Айзек Азимов. Сами боги / Н. Рыбакова]

Наконец, на материале многоязычного параллельного корпуса возможно типологическое сопоставление перфектов в языках Европы методом измерения расстояний (в программе NeighbourNet/SplitsTree, первоначально применявшейся в биологических исследованиях [Huson, Bryant 2006]; ср. также аналогичное исследование перфекта в [Dahl 2014] и исследование императива в славянских языках в работе [von Waldenfels 2014]). Строится матрица, строки которой соответствуют языкам, а столбцы — предикациям текста; единица означает, что в этой точке встретился перфект, ноль — что употреблено другое средство выражения. На базе этой матрицы программа порождает граф рас-

стояний между рядами данных. Оказывается, что получившаяся сеть расстояний «группирует», помимо объединений, примерно изоморфных генетическим и ареальным (итальянский + испанский, нидерландский + немецкий), также вместе балканские языки и португальский язык, в которых перфект сохраняется прежде всего в контексте экспериенциального значения (материал — переводы «Алисы в стране чудес»):



Аналогичная картинка по переводам «Винни-Пуха» (где во включенном в корпус материале отсутствовали романские языки и греческий, но присутствовали два балтийских) показывает явную типологическую близость малоупотребительного в этом переводе литовского перфекта не к германским языкам с «сильным перфектом», а к балканским, в то время как латышский перфект занимает скорее промежуточную позицию.



Таким образом, объём и проработанность параллельных корпусов НКРЯ уже достаточны для проведения достаточно сложных со-

поставительных и типологических исследований, как на лексическом, так и на грамматическом уровне.

Литература

Бунтман Н. В., Зализняк Анна А., Зацман И. М., Кружков М. Г., Лощилова Е. Ю., Сичинава Д. В. Информационные технологии корпусных исследований: принципы построения кросслингвистических баз данных, Информатика и её применения, 8:2 (2014), 98—110

Добровольский Д. О. Корпус параллельных текстов в исследовании культурно-специфичной лексики // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 383—401.

Зализняк Анна А., *Левонтина И. Б., Шмелев А. Д.* Константы и переменные русской языковой картины мира. М.: Языки славянских культур, 2012.

Павлова А. В. (сост.) От лингвистики к мифу: лингвистическая культурология в поисках «этнической ментальности». СПб.: Антология, 2013.

Руссо М. М. Неогумбольдтианская лингвистика и рамки «языковой картины мира» // Политическая лингвистика, 1 (47), 2014, с. 12—24 (http://journals.uspu.ru/attachments/article/622/%D0%9F%D0%BE%D0%BB%D0%B8%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F%20%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_2014_1_%D1%81%D1%82.%2001.pdf)

Сичинава Д. В. Поливариантные параллельные тексты в составе НКРЯ // Труды международной конференции «Корпусная лингвистика-2013», СПб., 2013 (http://corpora.phil.spbu.ru/Works2013/%D0%A1%D0%B8%D1%87%D0%B8%D0%BD%D0%B0%D0%B2%D0%B0.pdf)

Сичинава Д. В. Европейский перфект сквозь призму параллельного корпуса // Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН / Отв. ред. Н. Н. Казанский. Т. XII. Ч. 2. Исследования по теории грамматики. Выпуск 7: Типология перфекта / Отв. ред. Т. А. Майсак, В. А. Плунгян, Кс. П. Семенова. СПб.: Наука, 2016.

Січінава Д. В., Тищенко-Монастирська О. О., Шведова М. О. Паралельні українсько-російський та російсько-український корпуси //Лексикографічний бюлетень, 20. К., 2011: 35–38

Cysouw M., Wälchli B. (eds.). Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in Sprachtypologie & Universalienforschung STUF 60.2, 2007.

Dahl Ö., Hedin E. Current relevance and event reference. // Osten Dahl (ed.), Tense and Aspect in the Languages of Europe, 385—402. Berlin, New York: de Gruyter, 2000.

Dahl Ö. The perfect map: Investigating the cross-linguistic distribution of TAME categories in a parallel corpus // Szmrecsanyi, Benedikt & Walchli, Bernhard. (eds.) Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech. Linguae & litterae 28. Berlin: Walter de Gruyter, 2014: 268–289.

Goldberg A. Constructions: A Construction Grammar Approach to Argument structure. Chicago, 1995.

Goldberg A. Constructions at Work. The nature of generalization in grammar. Oxford, 2006.

Hasselgård H., Oksefjell S. (eds.). Out of Corpora: Studies in. Honour of Stig Johansson. Amsterdam — Atlanta, GA:Rodopi, 1999.

Huson D. H., Bryant D. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23(2), 2006: 254—267.

Lindstedt J. The perfect — aspectual, temporal and evidential. // Östen Dahl (ed.), Tense and Aspect in the Languages of Europe, 365–384. Berlin, New York: de Gruyter, 2000.

Loiseau S., Sitchinava D. V., Zalizniak A. A., Zatsman I. M. Information technologies for creating the database of equivalent verbal forms in the Russian—French multivariant parallel corpus [Информационные технологии создания баз данных эквивалентных глагольных форм в русско-французском поливариантном параллельном корпусе] // Информатика и её применения, 7:2 (2013), 100–109.

McCoard R. W. The English Perfect: Tense-Choice and Pragmatic Inferences.Amsterdam: North-Holland, 1978.

Sitchinava D. Parallel corpora within the Russian National Corpus // Prace Filologiczne, LXIII, 2012, 271—278.

Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. Parallel

corpora for medium density languages // Proceedings of the RANLP 2005: 590-596.

v. Waldenfels R. Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment // Beitrage der Europaischen Slavistischen Linguistik (POLYSLAV) 9. Munchen, 2006, S. 123─138 (доступно по адресу http://www-nw.uni-regensburg.de/%7E.war05297.slavistik.sprachlit.uni-regensburg.de/pub/WaldenfelsParallelCorpora2006.pdf)

v. Waldenfels R. Explorations into variation across Slavic: Taking a bottom-up approach // Szmrecsanyi, Benedikt&Walchli, Bernhard. (eds.) 2014. Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech. Linguae & Litterae 28. Berlin: Walter de Gruyter, 290-323.

Dmitri V. Sitchinava
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
mitrius@gmail.com

PARALLEL TEXTS WITHIN THE RUSSIAN NATIONAL CORPUS: NEW DIRECTIONS AND RESULTS

The paper presents the current states of the parallel corpora within the RNC and the updates of the last six years. The parallel corpora with the RNC include the following bilingual Russian—X corpora: Armenian, Belarusian, Bulgarian, English, Estonian, French, German, Italian, Latvian, Polish, Spanish, and Ukrainian. Virtually for all these language we have language pairs in both directions of translation. The RNC now includes a multilingual corpus that consists of nine texts and involve more than 20 languages, mostly Slavic. The Russian-French corpus has been developed, since 2012, as a polyvariant corpus with multiple translations of the same text into the same language. Polyvariant texts are also included also in the multilingual corpus. Parallel corpora now include texts of different genres, not only fiction, but also news, technical, scientific, religious and

legal texts, viz. all the main subdivisions represented in the main RNC. The discrepancies between the original text and the translation are now specially marked (arbitrary omission, adding, or change by the translator). Most languages represented in the corpus also have a morphological tagset and annotation. Some examples of corpus-based studies of lexicon and grammar are examined in more detail.

Key words: Parallel corpus, annotation, multilingual corpus, lexical typology, grammatical typology, language-specific lexicon, perfect

References

Buntman N. V., Zaliznyak Anna A., Zatsman I. M., Kruzhkov M. G., Loshchilova E. Yu., Sichinava D. V. [Informational technologies in corpus research: framework for building cross-linguistic databases]. *Informatika i ee primeneniya*, 8:2 (2014), pp. 98–110 (In Russ.)

Cysouw, M., Wälchli B. (Eds.). *Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in Sprachtypologie & Universalienforschung STUF* 60.2, 2007.

Dahl Ö. The perfect map: Investigating the cross-linguistic distribution of TAME categories in a parallel corpus. Szmrecsanyi, Benedikt & Walchli, Bernhard. (eds.) *Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech.* Linguae & litterae 28. Berlin, Walter de Gruyter, 2014, pp. 268–289.

Dahl Ö., Hedin E. Current relevance and event reference. In: Osten Dahl (ed.), *Tense and Aspect in the Languages of Europe*, 385–402. Berlin, New York, de Gruyter, 2000.

Dobrovol'skij D. O. [Corpus of parallel texts in studying culture-specific lexicon]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus 2006–2008: New Results and perspectives]. St. Petersburg, Nestor-Istoria Publ., 2009, pp. 383–401. (In Russ.)

Goldberg A. *Constructions at Work. The nature of generalization in grammar.* Oxford, 2006.

Goldberg A. *Constructions: A Construction Grammar Approach to Argument structure.* Chicago, 1995.

Hasselgård, H., Oksefjell, S. (Eds.). *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam – Atlanta, GA:Rodopi, 1999.

Huson D. H., Bryant D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2), 2006, pp. 254–267.

Lindstedt J. The perfect — aspectual, temporal and evidential. In: Östen Dahl (ed.), *Tense and Aspect in the Languages of Europe*, 365–384. Berlin, New York, de Gruyter, 2000.

Loiseau S., Sitchinava D. V., Zalizniak A. A., Zatsman I. M. Information technologies for creating the database of equivalent verbal forms in the Russian–French multivariant parallel corpus. *Informatika i ee primeneniya*, 7:2 (2013), pp. 100–109.

McCoard R. W. *The English Perfect: Tense-Choice and Pragmatic Inferences*. Amsterdam, North-Holland, 1978.

Pavlova A. V. (Ed.) *Ot lingvistiki k mifu: lingvisticheskaya kul'turologiya v poiskakh «etnicheskoi mental'nosti»* [From linguistics to myth: linguistic cultural studies searching "ethnic mentality"]. St. Petersburg, Antologiya Publ., 2013.

Russo M. M. [Neo-Humboldtian linguistics and the boundaries of the "language image of the world"]. *Politicheskaya lingvistika*, 1 (47), 2014, pp. 12—24 (Available at: http://journals.uspu.ru/attachments/article/622/%D0%9F%D0%BE%D0%BB%D0%B8%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F%20%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_2014_1_%D1%81%D1%82.%2001.pdf, accessed on 4.7.2015) (In Russ.)

Sichinava D. V. [Polyvariant parallel texts within the RNC]. *Trudy mezhdunarodnoi konferentsii «Korpusnaya lingvistika-2013»* [Proceedings of the international conference "Corpus linguistics 2013"]. St. Petersburg, Saint Petersburg University, 2013 (Avaliable at http://corpora.phil.spbu.ru/Works2013/%D0%A1%D0%B8%D1%87%D0%B8%D0%BD%D0%B0%D0%B2%D0%B0.pdf, accessed on 4.7.2015)

Sichinava D. V., Tyshchenko-Monastyrska O. O., Shvedova M. O. [Parallel Russian-English and English-Russian corpora]. *Leksikografichnyi byuleten'* 20, Kyiv, 2011, pp. 35–38. (In Ukrainian)

Sitchinava D. Parallel corpora within the Russian National Corpus. *Prace Filologiczne*, LXIII, 2012, pp. 271–278.

Sitchinava D. V. [European perfect within a parallel corpus] *Acta linguistica Petropolitana XII. Part 2. Studies in Grammar theory: Vol. 7: Typology of perfect.* St. Petersburg, Nauka Publ., 2016 [In Russ.].

Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, pp. 590–596.

v. Waldenfels R. Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment. *Beitrage der Europaischen Slavistischen Linguistik* (POLYSLAV) 9. Munchen, 2006, pp. 123–138 (available at http://www-nw.uni-regensburg.de/%7E.war05297.slavistik.sprachlit.uni-regensburg.de/pub/WaldenfelsParallelCorpora2006.pdf)

v. Waldenfels R. Explorations into variation across Slavic: Taking a bottom-up approach. Szmrecsanyi, Benedikt&Walchli, Bernhard. (eds.) 2014. *Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech*. Linguae & Litterae 28. Berlin, Walter de Gruyter, 2014, pp. 290–323.

Zaliznyak Anna A., Levontina I. B., Shmelev A. D. *Konstanty i peremennye russkoi yazykovoi kartiny mira* [Constants and variables of the Russian language image of the world]. Moscow, Languages of the Slavic cultures Publ., 2012.