Е.Г. – Интервьюируемая в мини-опросе (см. раздел 3), ж., 33 года, менеджер по персоналу большой американской международной компании в Москве

Е.И. – Интервьюируемая в мини-опросе (см. раздел 3), ж., 32 г., увольнение состоялось в 1999г., когда ей было 19 лет, в настоящее время менеджер в международном рекламном агентстве в Москве.

Е.П. – Интервьюируемая в мини-опросе (см. раздел 3), ж., 30 лет, руководитель службы персонала фабрики в Москве.

ИС 4 – Интервью автора <Р.Р.> с директором по персоналу крупной международной компании (ж., <Т.Я.>). М., март 2008 г.; продолжительность интервью: 70 мин.

ИС 9 – Интервью автора <Р.Р.> с генеральным директором мебельной фабрики(м., <В.К.>). Участвовали в разговоре также: педагог (м.) и научный сотрудник (ж.). Город С, май 2008 г. Продолжительность интервью: 100 мин.

СБ 9 – Собеседование с 9 претендентами на место директора филиала рекламного агенства; собеседование состоялось в офисе кадрового агентства в городе С в мае 2008 г. Участвуют: руководитель рекламного агентства (м., <Э.>), руководитель кадрового агентства (ж., <Е.>), секретарь кадрового агентства (ж., <О.>), кандидаты (все м.; <К1>, <К2>, … <К9>), австрийская профессор (ж., <Р.Р.>). Продолжительность записи: 130 мин.

СБ 10 – Собеседование с тремя претенденткими на место менеджера по обучению персонала; собеседование состоялось в городе С в мае 2008 г. Участвуют: директор по персоналу фирмы работодателя (ж., <С.К.>); ее заместитель (м., <А.Ш.>), менеджер по обучению персонала (ж., <С.>), менеджер по корпоративной культуре (ж., <Л.>), претендентки (ж., <К1>, <К2>, <К3>), австрийская профессор (ж., <Р.Р.>). Продолжительность записи: 50 мин.

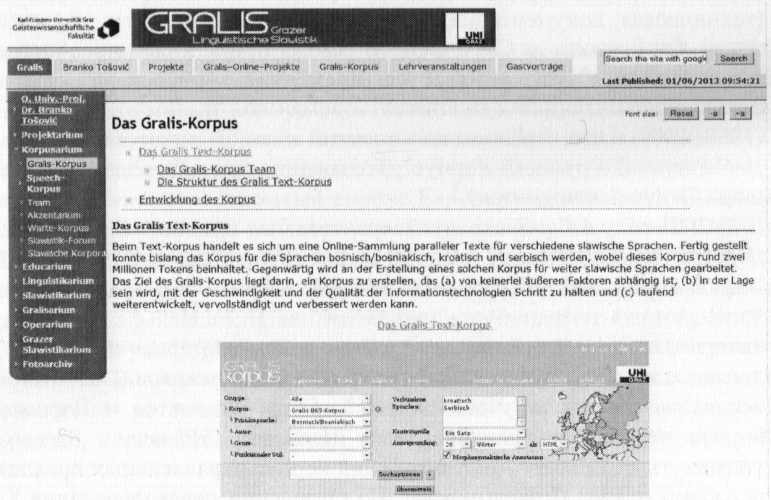
СБ 13 – Собеседование с претендентом на место инженера; собеседование состоялось в кадровом агентстве в городе С в мае 2008 г. Участвуют: менеджер по подбору персонала (ж., <М.>), кандидат (м., Р.>), австрийская профессор (ж., <Р.Р.>). Продолжительность записи: 20 мин.

Wiener Slawistischer Almanach, Sonderband 83 (2013), 89 - 111

Бранко Тошович

ГРАЛИС-КОРПУС

0. В Университете им. Карла и Франца в Граце (Австрия; Karl-Franzens-Universität Graz) в 2007 году создан комплексный и многоязычный текстуальный и устный корпус для изучения славянских языков и обучение им.[1] Корпус в настоящее время используется славистами из различных стран в рамках коллективных и индивидуальных проектов и в процессе обучения.[2] Его можно использовать как многоязычный (параллельный), а



также как одноязычный корпус, но в первую очередь предназначен для параллелизации языков генетически близких (славянских со славянскими) и неродственных (славянских с германскими). Типологически Гралис-

[1] http://www-gewi.kfunigraz.ac.at/gralis/index.html.

[2] Руководителем проекта является Бранко Тошович (Branko Tošović), координатором работ – Арно Вониш (Arno Wonisch), работу над развитием программной поддержки возглавляет Хуберт Штиглер (Hubert Stigler), а администратором корпуса является Ольга Ленер (Olga Lehner).

Корпус относится к аннотированным корпусам (морфосинтаксическая разметка пока сделана для сербского, хорватского, бошняцкого и черногорского языков). Пользователи имеют в распоряжении оболочку (интерфейс) на английском, немецком, русском и сербском/хорватском/бошняцком (число языков постепенно увеличивается).

Параллельные корпусы

1. Существует мало параллельных корпусов для славянских языков (далее СлЯз). Их нет так много и для неславянских языков. Среди русско-неславянских корпусов выделяются те, которые разработаны в рамках „Национального корпуса русского языка“ и которые охватывают восемь языков: английский, белорусский, испанский, итальянский, немецкий, польский, украинский, французский (Ruscorpora-www),[3] среди которых один – Русско-немецкий и немецко-русский корпус (автор Сергей Шаров) размещен в Лидсе (Leeds Corpus). В рамках проекта „Opus“ (an open source parallel corpus) подготовлен набор свободно распространяемых параллельных текстов (техническая документация, корпус субтитров) на пяти языках (Opus-www). Он основан на CWB CQP. В рамках проекта „Корпус Австрийской Академии наук“ готовится „Русско-немецкий корпус параллельных текстов“, охватывающий один текст – роман Ф. М. Достоевского „Идиот“ (1868–1869) и его переводы на немецкий язык.[4] В процессе создания находится немецко-русский корпус „Толкование снов Фрейда“ – „Traumdeutung“ (Traumdeutung-www).[5]

В 2001 году в Самарском университете был организован проект на тему „Корпусная лингвистика и новые информационные технологии“ в рамках межкафедрального сотрудничества с Институтом немецкого языкознания Университета г. Вюрцбурга (руководитель Н. Р. Вольф). Целью проекта является последовательное составление и компьютерная обработка параллельного немецко-русского корпуса переводных текстов (DER-Korpus) для исследовательских и учебных целей. Сюда относится и Параллельный корпус переводов „Слова о полку Игореве“ (SPI-www). Здесь можно упомянуть и „Lilabar“ – англо-русский корпус параллельных предложений, пословиц и фраз (Lilabar-www).[6] На Отделении переводоведения Тампер-

[3] Полная информация об интернет-источнике находится в списке литературы в конце работы.

[4] Этот корпус ориентирован на изучение лексической семантики в сопоставительном аспекте.

[5] Целью данного проекта является создание и морфологическое аннотирование немецко-русского корпуса для различных пользователей.

[6] Он содержит базу параллельных предложений („память переводов“, англ. translation memory) с возможностью поиска по ней. В базе представлены переводы английских и

ского университета (Финляндия) создан русско-финский параллельный корпус художественных текстов „ПарРус“.[7] К отдельной группе относится параллельный корпус, состоящий, с одной стороны, из текстов на английском и немецком языках, а с другой – из текстов на славянских языках (Bg, Be, Bs, Cs, Hr, Pl, Ru, Sk, Sr, Uk). Это The Regensburg Parallel Corpus (RPC) в Институте славистики Университета в Регенсбурге (RPC-www) [8], который переименован в ParaSol (PARAllel corpus of Slavic and Other Languages) и в настоящее время является совместным проектом Института славистики в Берне и Университета Регенсбург. Корпус сопровождает морфосинтаксическая аннотация. В этом университете разработан еще один корпус – Bilingualkorpus: Das Regensburger Korpus slavisch-deutscher Bilingualer (ReBiSlav:). В его основе находятся семинары с упражнениями (Hauptseminaren mit Übung) и одна магистерская работа. Он содержит ряд интервью со славянско-немецкими двуязычными говорящими различного уровня (сами студенты делают записи и транскибируют их). Сюда также относится „Восточный многоязычный корпус“, охватывающий болгарский, чешский, эстонский, венгерский, румынский и словенский языки.

Славяно-славянских параллельных корпусов не так много. Здесь можно упомянуть „Русско-словацкий параллельный корпус“ и „Украино-русский параллельный текстовый корпус“, состоящий из веб-публикаций, доступных для поиска в режиме онлайн и для скачивания (Infostream-www).

Среди неславянских параллельных корпусов выделяется Europarl Parallel Corpus (EPC-www) – открытый корпус Европарламента на 11 языках, в рамках которого развиты подкорпусы Bulgarian-English, Czech-English, Danish-English, German-English, Greek-English, Spanish-English, Estonian-English, Finnish-English, French-English, Hungarian-English, Italian-English, Lithuanian-English, Dutch-English, Polish-English, Portuguese-English, Romanian-English, Slovak-English, Slovene-English, Swedish-English. Центр теории перевода Университета Leeds развил „Leeds Corpus“ (Leeds-www), охватывающий различные языки (английский[9], арабский, греческий, итальянский, испанский, китайский, немецкий, польский, португальский,

русских предложений (фраз) в обоих направлениях. Предложения (фразы) разделены по темам.

[7] Для проверки данных использовались относительно небольшие по объему англо-русский и русско-английский массивы художественных текстов. В его состав входят произведения русской литературы XIX–XX вв. и их переводы на финский язык.

[8] Он охватывает только один функциональный стиль – литературно-художественный (тексты в подлиннике и в переводе).

[9] Английский корпус состоит из новостей агентства Reuters. Использование корпуса ограничено только в целях исследования.

русский[10], финский, французский, японский). Одним из параллельных корпусов является „MAASTR“ (Maastr-www), охватывающий тексты Маастрихского договора на немецком и английском языках. В рамках немецко-французского проекта „Коллокации в контексте“ создан параллельный корпус „Немецко-французский словарь коллокаций“, включающий немецкие тексты с французским переводом и французские тексты с немецким переводом (Kollokation-www).[11] Это по сути словарь сочетаемостей, охватывающих типичное и постоянное окружение какого-либо лексического элемента, в первую очередь прилагательных и существительных. К этой группе примыкает также „Parallel Corpus of Portuguese and English“, сокр. COMPARA (Compara-www).[12] В 1999 году работа началась в Институте немецкого языка в Мангейме (Institut für Deutsche Sprache, IDS) работа над проектом „GeFrePac“ (German-French Reciprocal Parallel Corpus), финансируемым ELRA (European Language Resources Agency, Paris) и IDS, под руководством Вольфганга Тойберта (Wolfgang Teubert). Сюда относятся и другие корпусы (некоторые из которых являются лишь попытками), в основном ориентированные на английский язык: „Английско-германский корпус по переводу“ – „Chemnitz German-EnglischTranslation Corpus“, Английско-чешский корпус „Kacenka“, Англо-китайский параллельный корпус „HKUST“, созданный в Гонконге, Англо-норвежский параллельный корпус, Английско-французско-испанский корпус „Lancaster's ITU“, Англо-хорватский параллельный корпус, „Intersect“ – англо-немецкий параллельный корпус, „Agenda 21“, включающий датский, английский, французский и немецкий языки, „Многоязычный корпус“, содержащий переводы Библии на английском, вьетнамском, греческом, датском, испанском, латинском, финском, французском, шведском, языках и т. д.

2. Создание Гралис-корпуса было осуществлено в рамках **(а)** научно-исследовательского проекта (2006–2010), финансируемого Австрийским фондом для поддержки научных исследований (FWF: Fonds zur Förderung der wissenschaftlichen Forschung: FWF-Projekt, P19158-G03), **(б)** сотрудничества между Институтом славистики и Центром моделирования информации Гуманитарного факультета Университета им. Карла и Франца в Граце (Zentrum für Informationsmodellierung in den Geisteswissenschaften an der Karl-Franzens-Universität Graz) и **(в)** при поддержке специалистов в области славянской корпусной лингвистики.[13] В качестве платформы для корпуса используется Gralis – славистический лингвистический портал Университета им. Карла и Франца в Граце)[14], который в 2011 году отметил десятилетний юбилей.

[10] Он содержит лишь тексты новостей из „Известий“ в период 2000–2001 г. (объем 14 564 884 словоупотреблений). Здесь использован и русский „Referenzkorpus“ (50 512 584 словоупотреблений).

[11] Он состоит из CELEX-документов (право Европейского Содружества – соглашения, внешние отношения, законы) и документов Европейского парламента (EUROPARL).

[12] При его создании особенно подчеркивалась проблема авторских прав и указывалось на то, что тексты, авторами которых являются умершие 70 или больше лет назад, не нуждаются в получении разрешения.

В данном проекте преследуются следующие цели: создать параллельный корпус для всех СлЯз, пополнять его сбалансированным материалом, отражающим языковое разнообразие, проводить лингвистический, социолингвистический и психолингвистический[15] анализ данных корпуса в рамках совместных и индивидуальных проектов, а также отдельных исследований.

3. Гралис-Корпус развивается в первую очередь для того, чтобы ответить на вопрос, в чем отличаются славянские литературные языки (СлЯз) между собой и насколько они далеки друг от друга. В нем представлены все СлЯз (14): белорусский (Be), болгарский (Bg), боснийский/бошняцкий (Bs), лужицко-сербский (Ls),[16] македонский (Mk), польский (Pl), русский (Ru), сербский (Sr), словенский (Sl), словацкий (Sk), украинский (Uk), хорватский (Hr), черногорский (Mo), чешский (Cs), а также немецкий (De). По мере необходимости будут включаться и так наз. славянские микроязыки: градишчанско-хорватский (Hg), кашубский (Ks) и русинский (Rs).[17]

Электронные корпусы и близость/отдаленность языков

4. О степени близости/отдаленности СлЯз существуют различные мнения. Они особенно поляризованы по отношению к очень близким СлЯз, какими в частности являются Bs, Hr, Mo и SR или же Bg и Mk. Противоположные взгляды особенно проявляются в социолингвистическом определении характера СлЯз – являются ли они (а) отдельными языками, (б) вариантами одного (полицентрического) языка, (в) диалектами одного национального языка, (г) одним языком с различными (политическими) названиями и т. п. Отсутствие исследований, которые бы на базе электронных корпусов объективно представили соотношения СлЯз и предоставили релевантный материал для разъяснений всех существенных вопросов, создает благопри-

[13] В этих целях был проведен в летнем семестре 2006 года семинар на тему „Славянская корпусная лингвистика“.

[14] http://www-gewi.uni-graz.at/gralis.

[15] Имеется в виду восприятие языков носителями и неносителями языков, особенно близкородственных.

[16] Система Гралис-корпуса позволяет вносить тексты для обоих лужицкосербских языков.

[17] Латинские сокращения, указывающие на название языков, используются нами во всех работах, написанных на различных языках.

ятные условия для субъективных, тенденциозных толкований взаимодействия СлЯз и политизации их межъязыковых корреляций, во что втягиваются не только языковеды, но очень часто и нефилологи (в первую очередь политики). Данная проблема сильно обострилась после распада трех славянских федеративных государств (СССР, Чехословакии и Югославии) в конце XX столетия и военными конфликтами в юго-восточной Европе в это время. Ситуация особенно осложнилась из-за различных подходов к оправданности и неоправданности официального провозглашения и кодификации некоторых СлЯз в XX столетии (в середине этого века MK, а в 90-е годы Bs, Hr, Mk, Mo и Sr). На усложнение межславянских языковых отношений повлияло и изменение статуса отдельных СлЯз. Так (а) Cs и Sk (б) Be и Uk, (в) Sl превратились из языков республик в языки новых государств. С другой стороны, усилились тенденции к вытеснению Ru как языка общенационального общения в Белоруссии и Украине, в которой украинское и русскоязычное население втянуто в конфликтную ситуацию. Статус языка взаимного общения на уровне одного государства потерял и сербохорватский язык, в рамках которого после распада бывшей Югославии кодифицированы три (Bs, Hr, SR), а четвертый находится в этом процессе (Mo, провозглашенный совсем недавно – в 2007 году). Некоторые из представителей так наз. микрославянских языков требуют в новых государствах другого статуса (карпатско-русинский в Украине). Во всем этом появилась еще одна новизна – с распадом общего государства некоторые СлЯз стали языками национальных меньшинств (например, Hr в Сербии или SR во Хорватии). Определенная, а порой и значительная часть русскоговорящих оказалась за пределами Российской Федерации и при расширении Европейского содружества вошла в его состав (Латвия, Литва, Эстония). Почти все спорные вопросы так или иначе связаны с процессами объединения в рамках Европейского Содружества, которому приходится считаться со сложными межславянскими языковыми отношениями и искать решения. Непростая и комплексная языковая ситуация на территории СлЯз требует объективных исследований интралингвистических, социолингвистических и психолингвистических аспектов их соотношения на репрезентативном корпусном материале. Тот факт, что De находится в прямом или посредственном контакте со всеми СлЯз и что немецкоязычные страны имеют разветвленное сотрудничество со славянским миром, создает необходимость включить в корпус и славяно-немецкий языковой материал. Все вышеуказанные причины побудили нас создать корпус для системнолингвистического, социолинвгистического и психолингвистического изучения (в первую очередь) отношений между (а) всеми СлЯз и (б) СлЯз и De.

На материале Гралис-Корпуса можно будет рассматривать и вопрос о том, как носителями СлЯз воспринимается и оценивается расстояние между СлЯз, причем в центре внимания находится взаимное понимание, а также кодовое переключение (code-switching).[18] Данный корпус является важным для славистики, так как можно на однообразном и унифицированном материале проводить системное и комплексное исследование отношений между СлЯз, а также их параллели с De. Результаты корпусных работ укажут на индивидуальную межъязыковую славянскую и неславянскую (славяно-германскую) близость/отдаленность. Гралис-Корпус может послужить хорошей материальной основой для дальнейшего исследования СлЯз и De в их взаимных связях. Для германистики он даст полезный материал для определения расстояния между De и СлЯз, языками, относящимися к различным генетическим семьям. В настоящее время в составе ЕС находится пять СлЯз: Bg, CS, Pl, Sk и Sl, в ближайшее время такой статус получит Hr, на что также претендуют BS, Mo, Mk и Sr, т. е. десять СлЯз могут в ближайшие годы оказаться в ЕС. Одной из проблем является то, что значительная часть языков стран, рассчитывающих на членство в ЕС, относится к группе близкородственных, носители которых свободно, без переводчика общаются друг с другом (типичный пример – бошняки, сербы, хорваты, черногорцы), поэтому в рамках ЕС ведутся дискуссии о том, что делать с такими языками и надо ли для них создавать службу перевода. Создаваемый корпус может дать объективную картину о различиях между ними и о том, насколько оправдано, целесообразно это делать. Гралис-Корпус предоставит драгоценный материал тем, которые занимаются переводом, так как в его основе лежат „битексты“[19] и точки

[18] Некоторые исследователи считают его основным критерием для различения языков (если говорящие друг друга не понимают, они говорят на различных языках, если же понимают друг друга, речь идет об одном языке или диалекте). Хадсон приводит причины, почему к критерию взаимного понимания надо осторожно относиться, в частности: существуют варианты, которые можно считать различными языками, но их носители понимают друг друга (скандинавские языки, кроме финского и лапонского), но и у вариантов, принадлежащих к одному языку, может отсутствовать взаимное понимание (диалекты китайского языка) (Hudson 1990, 35–37). Анализируя тесты понимания как посредственного метода измерения лингвистического расстояния, Амон пришел к выводу, что отмеченную в этих тестах степень взаимного понимания нельзя считать однозначным индикатором лингвистического расстояния (Ammon 1986, 13). Более подробно об измерении лингвистического расстояния и лингвистического несходства, а также о тестах по измерению взаимного (не)понимания Амон писал в другой работе (Ammon 1989, 31–46).

[19] Битекст представляет собой параллельный текст на одном языке вместе с его переводом на другой язык. Это название ввел Брайан Хэррис (Brian Harris) в 1988 году, а концепция, базирующаяся на нем, была развита группой ученых при Университете Монреаля (Université de Montréal) под названием RALI *(Recherche appliquée en*

соприкосновения с концепцией памяти переводов.[20] Для студентов он особенно полезен, из-за наличия различных баз данных, вытекающих из него или связанных с ним. Кроме того, студентам предоставляется возможность собирать и исследовать материал для курсовых, дипломных, магистерских, кандидатских и докторских работ.

Когда Гралис-Корпус будет полностью готов, при его помощи можно будет искать ответ на вопрос, насколько близки/далеки языки, относящиеся к генетически одной языковой группе (славянской), и языки, принадлежащие к различным языковым семьям (германской и славянской), а также ответить на вопросы: (1) какие лингвистические, социолингвистические и психолингвистические процессы влияют на близость/отдаленность СлЯз, (2) насколько лингвистическое расстояние между СлЯз воздействует на языковую ситуацию и психолингвистические процессы, в первую очередь на порог узнавания и понимания, (3) насколько СлЯз приближаются или удаляются друг от друга в четырех различных группах: в группе очень близких языков (например, Bg и Mk), в группе территориально близких языков – восточнославянских, западнославянских и южнославянских (например, Be, Ru и Uk) и в группе территориально отдаленных языков (например, HR и Cs).

linguistique informatique или *Applied Research in Computational Linguistics* – «Прикладные исследования в вычислительной лингвистике»). „В сфере исследований в области перевода «битекст» – это совмещенный документ, состоящий из версий соответствующего текста на исходном и целевом языках. Битексты создаются с помощью специальных компьютерных программ, которые называются «инструментами для выравнивания» (*alignment tool*) или «инструментами для битекста» (*bitext tool*), которые позволяют автоматически выравнивать оригинальную версию текста и его перевод. Подобные программы, как правило, приводят в соответствие два текста (оригинал и перевод) по каждому предложению. Собрание битекстов называется «битекстовой базой данных» или «двуязычным корпусом» и может использоваться в качестве справочника и для поиска нужных сочетаний“ (http://ru.wikipedia.org/ wiki/ Битекст).

[20] Различие между битекстом и памятью переводов состоит в том, что память переводов представляет собой базу данных, в которой сегменты текста (соответствующие друг другу предложения) не связаны с оригинальным контекстом (оригинальная последовательность предложений теряется): „Битекст же сохраняет изначальную последовательность предложений. Стандартным форматом для обмена базами данных памяти переводов между разными программами автоматизированного перевода является формат TMX (XML словарь, опубликованный LISA (Ассоциация отрасли локализации – Localisation Industries Association). TMX позволяет сохранять оригинальный порядок предложений. Битексты создаются в качестве справочного инструмента для консультаций специалистов-переводчиков, а не автоматизированных программ. Поэтому небольшие ошибки выравнивания или неточности, которые могут привести к сбоям в памяти переводов, для них не имеют значения“ (http://ru.wikipedia.org/wiki/ Битекст).

Целью корпуса является предоставление разнообразного материала для изучения совпадений, сходств и различий между СлЯз. Так как речь идет о близких языках и так как их отношение служит поводом для разнообразных спекуляций, данный корпус может послужить материальной основой для объективной оценки. Гралис-Корпус преследует цель показать, как СлЯз функционируют и взаимодействуют на конкретном материале, в реальном контексте и на всех языковых уровнях (фонетико-фонологическом, орфоэпическом, грамматическом и стилистическом). В подготовке Гралис-Корпус делается упор на то, чтобы он был как можно более репрезентативным и сбалансированным. Гралис-Корпус готовится в рамках концепции языковых корреляций, изложенной в книге *Korelaciona sintaksa* (Tošović 2001) и в ряде статей, особенно в работах, в которых рассматривается теоретический вопрос о том, что такое различие как понятие и различие в языке (Tošović 2008e).[21] В качестве теоретической основы для типологической обработки и представления текстов служит книга *Funkcionalni stilovi* (Tošović 2002). Согласно концепции, изложенной в ней, Гралис-Корпус разделен на пять функциональных стилей (литературно-художественный, публицистический, научный, официально-деловой и разговорный). При создании корпуса учитывались следующие критерии: (1) он не должен зависеть от любых внешних обстоятельств, (2) он должен быть в состоянии следить за ходом и скоростью изменений в области информационных технологий и (3) иметь возможность постоянно развиваться, совершенствоваться и дорабатываться. Так как качество любого корпуса определяется (а) глубиной и широтой аннотирования, (б) возможностями поиска и получения информации, (в) представительным (репрезентативным) характером, пропорциональностью и сбалансированностью, а также (г) свободой доступа, всему этому уделяется большое внимание. Гралис-Корпус приспособлен в значительной степени к программным мировым стандартам (TEI, XML и др.). Его развитие направлено на количественное расширение новым содержанием, качественное улучшение оболочки, более широкое и глубокое аннотирование, функциональное ускорение, улучшение поисковой системы и программное обновление.

Корпус предназначен в первую очередь для специалистов в области сопоставительного языкознания и для языковедов более широкого профиля (особенно в области общего и системного языкознания, а также социолингвистики), для интересующихся интралинвистическими, интерлингвистическими и экстралинвистическими отношениями между СлЯз. Он является полезным для всех тех, которые сталкиваются с проблемой

[21] Некоторые теоретические и практические аспекты отношений между близкородственными языками рассмотрены и описаны в сборнике *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen* (Tošović 2008).

взаимодействия СлЯз. Его можно целесообразно использовать в обучении, особенно в вузах.

Концептуальная основа Гралис-Корпуса

5. В основе стратегии развития Гралис-Корпуса лежит наша концепция славянских корреляционных систем, состоящая из интракорреляционала, интеркорреляционала, супракорреляционала, суперкорреляционала и экстракорреляционала. Интракорреляционал представляет собой сеть отношений внутри одного языка, влияющих на межъязыковом уровне (например, изменения внутри Ru, влияющие на расстояние межу Ru и Pl). Интеркорреляционал образуют очень близкие языки, языки с очень высокой степенью понимания (скажем, BS, Hr, Mo, Sr). Супракорреляционал состоит из территориально близких СлЯз (восточнославянских, западнославянских, южнославянских). Суперкорреляционал охватывает языки, относящиеся к территориально различным славянским группам (напр. восточнославянским и южнославянским типа Ru ↔ Bg). Экстракоореляционал является системой отношений генетически различных языков, в данном случае славянских языков с одной стороны и немецкого с другой.

Первая корреляционная система – интракореляционал состоит из отношений в рамках только одного языка. Предметом исследования данного типа при помощи корпусного материала являются **(а)** динамические процессы, а именно изменения в отдельных языках, влияющие на межъязыковое расстояние и ведущие к увеличению или сокращению расстояния между СлЯз, к (не)восприятию нововведений носителями определенного языка, к положительному или отрицательному восприятию, к усилению или ослабеванию понимания носителями других славянских языков, т. е. к увеличению или понижению порога узнавания и понимания, **(б)** статические процессы, а именно как и насколько сложившиеся структурно-типологические свойства воздействуют на межъязыковые связи. Для каждого языка в интракорреляционале существует главная временная линия разграничения, чаще всего важное для языковой проблематики событие: например, для Be, Ru и Uk это распад СССР (1989), для Cs и Sk – распад ЧССР (1993), для Bs, Hr, Mo и Sr – распад СФРЮ (1992), а для De – объединение Германии (1990). Если нет отчетливой точки отсчета, то можно учитывать временной ориентир для других языков данной группы (например, для Pl и Ls в группе западнославянских языков – 1993 год, для Bg в группе южнославянских языков – 1992). Для определения интракорреляционного расстояния между СлЯз при помощи такого корпуса необходимо иметь, как минимум, две версии из различных периодов.

Вторая корреляционная система – интеркореляционал охватывает очень близкие СлЯз, какими являются (а) BS, Hr, Mo, Sr, (б) Bg и Mk, (б) Sk и Cs. Проведенное исследование (Tošović 2008) свидетельствует о том, что процесс их дивергенции усилился после распада бывшей Югославии, что повлияло на увеличение расстояния между BS, Hr, Mo, Sr. Этому особенно способствуют радикальные пуристические тенденции в некоторых из них и обострение общественных процессов (усиление национализма и шовинизма, развязывание военных конфликтов и т. п.) и психические факторы (напр., ненависть к другим народам и их языкам). Целью корпусных работ в данной части является не только подготовка корпусного материала для изучения отношений между СлЯз внутри каждого интеркорреляционала (**A**, **B** и **C**), но и вопрос о межгрупповом расстоянии (скажем, является ли расстояние между Hr и Sr бóльшим, чем между Bg и M или между Cs и Sk).

Супракорреляционал состоит из отношений территориально близких СлЯз. В его состав входит восточно-западнославянский супракореляционал: Be, Ru, Uk (**A**), западно-южнославянский супракореляционал: Cs, Ls, Pl, Sk (**B**) и южнославянский супракореляционал: Bg, Bs, Hr, Mk, Mo, Sr (**C**). Цель корпусной работы в рамках супракорреляционала – подготовка материала для определения различий между языками, входящими в состав каждого супракорреляционала в отдельности (скажем, расстояния между Ls, Pl, Sk и Cs), а также определение расстояния между языками, принадлежащими различным супракорреляционалам (здесь ставится вопрос, являются ли более близкими или более далекими языки супракорреляционала **A** по отношению к **B** или **C**, напр. Bg, Mk ↔ Cs, Sk, Cs, Pl ↔ Ru, Uk, Be, Ru ↔ Mk, Sl).

В составе суперкореляционала находятся СлЯз из различных групп – 1. восточно-западнославянской: а) Be, Ru, Uk, б) Ls, Pl, Sk, Cs, 2. восточно-южнославянской: а) Be, Ru, Uk, б) Bg, Bs, Hr, Mk, Mo, Sr), 3. западно-южнославянской: а) Cs, Ls, Pl, Sk, б) Bg, Bs, Hr, Mk, Mo, Sr. Одной из корпусных проблем в данной сети взаимодействий является вопрос, насколько интеркорреляционное изменение влияет на суперкорреляционное расстояние. Здесь будет предпринята попытка собрать корпусный материал для рассмотрения положения о том, что увеличение интеркорреляционного расстояния влияет на характер суперкорреляционного расстояния. Скажем, тяготеют ли некоторые процессы в Hr, направленные на сознательное удаление от Sr (= увеличение интеркорреляционного расстояния), к сближению Hr и Ru (= сокращение супракорреляционного расстояния).

Для определения интер-, супра- и суперкореляционного взаимодействия необходимы тексты, переведенные по возможности на все СлЯз или на большинство из них.

Особую систему составляют соотношения СлЯз с немецким языком в рамках экстракореляционала. Здесь в центре внимания находятся связи между этим языком и славянскими интракорреляционными языками (например, Bs, Hr, Mo, Sr), супракорреляционными языками (например, Cs, Ls, Pl, Sk) и суперкорреляционными языками (например, Sl и Uk, Pl и Be, Mk и Cs). Данная часть корпусных работ преследует цель собрать материал, чтобы ответить на вопрос, к каким СлЯз стоит ближе De, а от каких он находится подальше. Здесь надо проверить гипотезу о том, что наличие прямого территориального контакта влияет на сокращение расстояния между СлЯз и De. Для определения экстракорреляционного расстояния необходимо выбрать (а) славянские тексты, имеющие больше всего переводов на немецкий язык (по возможности уже включенные в корпус по определению интер-, супра- и суперкореляционного расстояния), и (б) немецкие тексты, имеющие больше всего переводов на СлЯз.

Каждая из вышеупомянутых корреляционных систем должна иметь свой подкорпус – Интра-Кор (только для одного языка), Интер-Кор (для очень близких СлЯз), Супра-Кор (для языков, относящихся к одной из трех групп СлЯз), Супер-Кор (для языков, относящихся к различным славянским группам), Экстра-Кор (для СлЯз и немецкого), и отличаться формальной унификацией и функциональной сбалансированностью.

Для исследования межславянских языковых отношений в их корреляционных системах в корпус надо включать переводы, опубликованные в новейшее время. При этом нужно учитывать то, какой из текстов является самым переводимым на СлЯз. В рамках литературно-художественного стиля основным материалом являются переводы прозаических произведений. Работа над созданием корпуса для публицистического стиля подразумевает поиск текстов с одинаковым, немодифицированным содержанием из on-line изданий типа „Southeast European Times“, „Deutsche Welle“, „Voice of America“ и т. п. Так как они покрывают лишь некоторые СлЯз, для остальных приходится делать переводы с уже имеющих языковых версий. При поиске материала для научного стиля упор делается на переводы собственно научных произведений. Для анализа отношений между СлЯз в рамках официально-делового стиля важнейшими являются славянские версии основных документов международных организаций (ООН, ЕС, ЮНЕСКО и др.). Что касается разговорного стиля, уже создается SlawSpeech-Korpus, и его материал используется для написания одной докторской диссертации. В его состав войдут три подкорпуса: SlawWort-Korpus, SlawFix-Korpus и SlawFrei-Kopus. Первые два будут использованы

для фонетического и просодического анализа. В SlawWort-Korpus будут вноситься записи отдельных слов, являющихся общими для всех СлЯз. Например, составляется список из двадцати слов, относящихся к различным частям речи и используемых во всех СлЯз. Эти слова читаются опрошенными. Запись потом разделяется на сегменты, которые соответствуют каждому слову, а потом они параллелизуются таким образом, чтобы при нажатии на одно слово появились все произнесенные версии на всех СлЯз. В SlawFix-Korpus войдут записи связанных слов в предложении. Такие тексты должны быть небольшими (не более двадцати предложений). Третий подкорпус – SlawFrei-Kopus предназначен для определения отношений между СлЯз на уровне текста и стиля. Он будет охватывать спонтанно произнесенные высказывания на одну из выбранных тем, что даст возможность измерять расстояние в речи, не подлежащей внешней языковой корректуре и вмешательству цензуры. Скажем, опрошенному дается рисунок или набор рисунков с просьбой рассказать, что там нарисовано. Так как данное исследование охватывает большое число языков, такие записи должны был короткими. В SlawWort-Korpus войдут слова, общие для СлЯз и De, а в SlawFix-Korpus (а) самые простые предложения на СлЯз для опрашиваемых с родным немецким языком и (б) самые простые предложения на De для опрашиваемых с родным СлЯз.

В процессе анализа корпусного материала проводятся следующие процедуры: **(а)** берется определенный материал (словарный или текстуальный) из двух различных периодов одного и того же языка (напр., 1970–1990, 1991–2010) и фиксируются изменения, которые влияют или могут влиять на расстояние (увеличение или сокращение) между СлЯз, **(б)** определяется характер этих изменений: почему они происходят (для более экономного и эффективного выражения, спонтанно или целенаправленно, с политическими целями, для усиления межъязыковой дивергенции или конвергенции и т. п.), являются ли они релевантными, случайными, спонтанными, запланированными, целенаправленными и др.

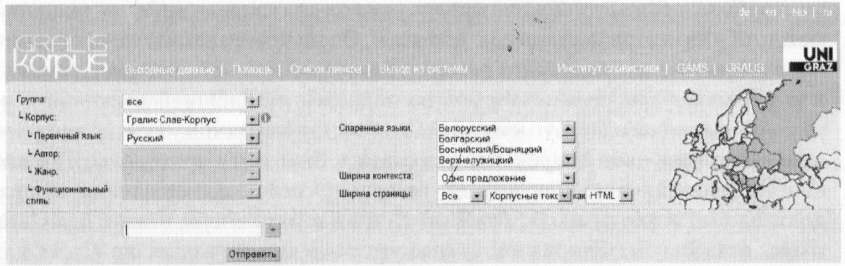
При заполнении корпуса материалом предусмотрено пропорциональное и уравновешенное включение текстов трех типов: оригинальных, модифицированных (адаптированных) и переведенных.

На собранном корпусном материале можно изучать влияние языковой политики, стандартизации и кодификации на сокращение или увеличение расстояния между СлЯз, а также рассматривать вопрос, как носителями СлЯз воспринимаются и оценивается расстояние между отдельными языками.[22]

[22] Здесь в центре внимания находится критерий взаимного понимания, а также влияния расстояния на кодовое переключение (code-switching).

Структура Гралис-Корпуса

6. Гралис-Корпус состоит из нескольких подкорпусов, разделенных на макрогруппы (восточнославянские, западнославянские и южнославянские языки) и микрогруппы (корпусы отдельных языков, а также индивидуальные корпусы – корпусы писателей, напр., Иво Андрича, Зорана Живковича и др.). Существует также возможность выбора двух макрогрупп (скажем, восточнославянской и южнославянской). Отдельный блок составляют славянские языки, соотнесенные с немецким языком.



7. В рамках Гралис-Корпуса выделяются две основные части: Корпус устной речи (Speech-Korpus) и Корпус письменных текстов (Text-Korpus).

8. Устный Гралис-Корпус (Gralis Speech-Korpus) охватывает транскрибированные записи и предоставляет возможность анализировать речевые единицы фонетически (акустически, артикуляционно), просодически и фонологически на уровне звука/фонемы, слога, лексемы, синтагмы, словосочетания и предложения. Существуют три подкопруса: Wort-Korpus, Fix-Korpus i Frei-Korpus. Wort-Korpus содержит записи отдельно произнесенных слов на разных славянских языков (в настоящее время больше всего на сербском, хорватском и бошняцком). Fix-Korpus охватывает записи озвученных текстов. Frei-Koprus состоит из записей спонтанной речи и их транскриптов.


Wort-Korpus

Speech Korpus

Wort balon suchen

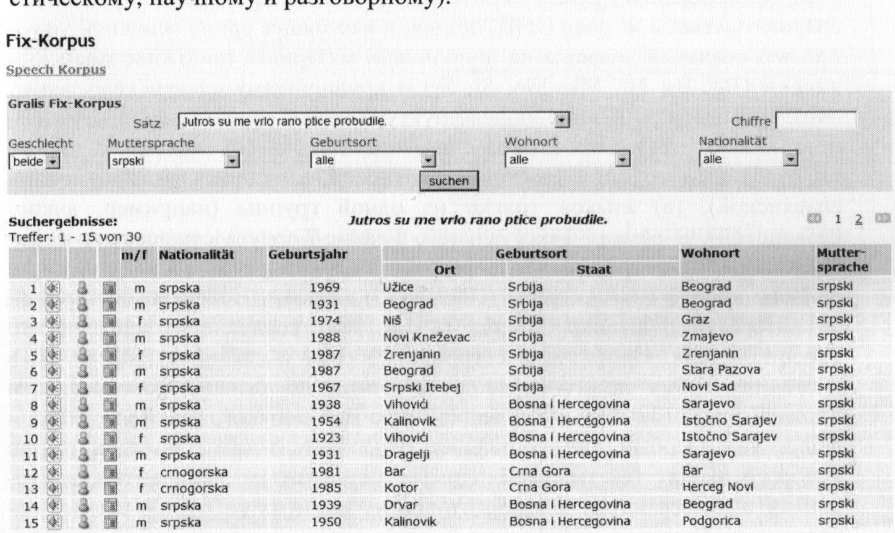
z.B. balon, ba*, m*

Suchergebnisse:
Treffer: 1 - 3 von 3

	Wort	m/f	Nationalität	Geburtsjahr	Geburtsort		Wohnort	Muttersprache
					Ort	Staat		
1	balon	f	hrvatska	1983	Zenica	Bosna i Hercegovina	Graz	hrvatski
2	balon	f	hrvatska	1980	Rotterdam	Holandija/Nizozemska	Osijek	hrvatski
3	balon	f	hrvatska	1977	Bjelovar	Hrvatska	Bjelovar	hrvatski

Фонетическая и просодическая транскрипция проходит верификацию экспертов (как минимум одного) при помощи специальной программы – Valorisarium'a. Speech-Korpus связан с программой Akzentarium, предоставляющей возможность получить информацию об ударении (в данный момент для более чем 120 000 слов сербского, хорватского и бошняцкого языков) и позволяющей ставить ударение и получать информацию о внесенных в нее словах. В Transkriptarium'е дается информация о том, как произносятся слова, а Suprasemgentarium указывает на интонационную структуру предложения.

Frei-Korpus является коллекцией свободно/спонтанно произнесенных текстов, относящихся ко всем функциональным стилям (литературно-художественному, публицистическому, официально-деловому, публицистическому, научному и разговорному).


Fix-Korpus

Speech Korpus

Gralis Fix-Korpus

Satz: Jutros su me vrlo rano ptice probudile. Chiffre

Geschlecht: beide | Muttersprache: srpski | Geburtsort: alle | Wohnort: alle | Nationalität: alle

suchen

Suchergebnisse: *Jutros su me vrlo rano ptice probudile.* 1 2

Treffer: 1 - 15 von 30

	m/f	Nationalität	Geburtsjahr	Geburtsort		Wohnort	Muttersprache
				Ort	Staat		
1	m	srpska	1969	Užice	Srbija	Beograd	srpski
2	m	srpska	1931	Beograd	Srbija	Beograd	srpski
3	f	srpska	1974	Niš	Srbija	Graz	srpski
4	m	srpska	1988	Novi Kneževac	Srbija	Zmajevo	srpski
5	f	srpska	1987	Zrenjanin	Srbija	Zrenjanin	srpski
6	m	srpska	1987	Beograd	Srbija	Stara Pazova	srpski
7	f	srpska	1967	Srpski Itebej	Srbija	Novi Sad	srpski
8	m	srpska	1938	Vihovići	Bosna i Hercegovina	Sarajevo	srpski
9	m	srpska	1954	Kalinovik	Bosna i Hercegovina	Istočno Sarajev	srpski
10	f	srpska	1923	Vihovići	Bosna i Hercegovina	Istočno Sarajev	srpski
11	m	srpska	1931	Dragelji	Bosna i Hercegovina	Sarajevo	srpski
12	f	crnogorska	1981	Bar	Crna Gora	Bar	srpski
13	f	crnogorska	1985	Kotor	Crna Gora	Herceg Novi	srpski
14	m	srpska	1939	Drvar	Bosna i Hercegovina	Beograd	srpski
15	m	srpska	1950	Kalinovik	Bosna i Hercegovina	Podgorica	srpski

В Speech-Korpus'е пользователю предоставляется несколько возможностей в поиске: по отмеченным предложениям, по частям с транскрипцией, акцентуацией и супрасегментацией и т. д.

Для обработки данных о записях используются особые процедуры, начинающиеся с заполнения бланка с основными персональными данными. Он разделен на три части: Опрошенный, Запись и Корпус. В первую графу вносится основная информация о говорящем (пол, национальность, религия, год и место рождения, место жительства, профессия, место обучения, родной язык, важнейшие места проживания, иностранный язык и т. п.), дата и место записи, во второй следуют данные о теме, месте, обстоятельствах записи, оборудовании, длине и формате записи, операторе, а в третью вносятся данные о том, для какого типа исследования использу-

ется записанный материал (для устного, спектрального или какого-нибудь другого анализа).[23] В конце бланка находится графа для письменного согласия говорящего на публичное представление (за определенным шифром). В этих целях выбирается пароль, состоящий из названия места записи, трехзначного номера и букв, обозначающий язык (напр., **b**, **k s** указывают на боснийский/бошняцкий, хорватский или сербский) типа split_001k. После внесения пароля открывается полный бланк с данными об опрошенном и о записи. Если нужно добавить дополнительную информацию, скажем, о месте или языке, нужно войти в графу „Внести новое место или пароль".

9. Текстуальный Гралис-Корпус (Gralis Text-Korpus) дает возможность спаривать тексты на всех СлЯз, причем в настоящее время основной уклон при его создании делается на пополнении материала по южнославянским языкам (Bg, Bs, Hr, Mk, Mo, Sl, Sr) и крупнейшему славянскому языку (Ru). Благодаря созданной инфраструктуре можно проводить параллелизацию (а) всех славянских языков и немецкого языка, (б) славянских языков по отдельным группам (восточнославянской, западнославянской, южнославянской), (в) языков только из одной группы (например, южнославянской). В рамках Текстуального Гралис-Корпуса существуют индивидуальные подкорпусы. Пока разработаны четыре из ряда запланированных: корпус Иво Андрича (1892–1975, лауреата Нобелевского премии за литературу), Бранко Чопича (1915–1984, величайшего славянского повествователя, юмориста и сатирика), Зорана Живковича (род. 1948, в настоящее время самого переводимого писателя бывшей Югославии) и Блаже Конеского (1921–1993, виднейшего македонского писателя, поэта и филолога). Особый тип составляет образовательный корпус, в состав которого входят тексты, необходимые дипломантам, аспирантам, докторантам и диссертантам для подготовки их научных работ. Со всеми подкорпусами Текстуальный Гралис-Корпус содержит около шести миллионов словоупотреблений. Для свободного использования он открыт в той части, в которой решен вопрос об авторских правах.

В Text-Korpus включаются тексты всех функциональных стилей и сопровождаются основной метаязыковой и грамматической аннотацией. Gralis-Korpus имеет три типа разметок: метатекстуальную, экстралингвистическую и морфосинтаксическую. Последняя и поиск по ее параметрам проводится при помощи особой программы – „MorphoGenerator".

[23] В анализе устного материала используются два метода: акустический (при помощи программы типа Praat измеряется расстояние между СлЯз по основным параметрам: длине произнесенных звуков, изменению в интонации и т. п.) и слуховой.

most		
Kasus	**Singular**	
	m	
Nominativ	most	N-msn----21Sm-08
Genitiv	mosta	N-msg----21Sm-08
Dativ	mostu	N-msd----21Sm-08
Akkusativ	most	N-msa--n-21Sm-08
Vokativ	moste	N-msv----21Sm-08
Instrumental	mostom	N-msi----21Sm-08
Lokativ	mostu	N-msl----21Sm-08

mosti mostovi		
Kasus	**Plural**	
	m	
Nominativ	mosti mostovi	N-mpn----21Sm-08 N-mpn----21Sm-08
Genitiv	mosta mostova	N-mpg----21Sm-08 N-mpg----21Sm-08
Dativ	mostima mostovima	N-mpd----21Sm-08 N-mpd----21Sm-08
Akkusativ	moste mostove	N-mpa----21Sm-08 N-mpa----21Sm-08
Vokativ	mosti mostovi	N-mpv----21Sm-08 N-mpv----21Sm-08
Instrumental	mostima mostovima	N-mpi----21Sm-08 N-mpi----21Sm-08
Lokativ	mostima mostovima	N-mpl----21Sm-08 N-mpl----21Sm-08

С текстуальным корпусом связан онлайн-словарь „Lexikarium" и программа „Akzentarium", позволяющие получать комплексную просодическую, лексикосемантическую и грамматическую информации.[24] Корпус предоставляет два типа поиска: простой и сложный (при помощи т. наз. CQP-синтаксиса). Результаты поиска отмечаются желтым или синим цветом. Над полученными предложениями стоит стрелочка, указывающая на подлинник (если речь идет об интернет-адресе, приводится ссылка). Из корпуса можно извлекать частотные списки.

Мост на Дрине	
Russisch	Там , где Дрина всей тяжестью своей зеленой и вспененной водной лавины извергается как бы из сомкнутой стены отвесных черных гор , стоит большой каменный мост строгих пропорций с одиннадцатью широкими пролетами .
Serbisch	Na tom mestu gde Drina izbija celom ležinom svoje vodene mase , zelene i zapenjene , iz prividno zatvorenog sklopa crnih i strmih planina , sloji veliki , skladno srezani mrsi od kamena , sa jedanaest lukova širokog raspona .
Deutsch	An dieser Stelle , wo die Drina mit dem ganzen Gewicht ihrer Wassermassen , grün und schäumend , aus dem scheinbar geschlossenen Massiv der schwarzen und steilen Berge hervorbricht , steht eine große , gleichmäßig geschnittene Brücke aus Stein mit elf weitgespannten Bögen .
Bulgarisch	На това място , където Дрина с цялата тежест на своите зелени и пенливи водни маси изблива от затворения наглед масив на черните и стръмни планини , има голям , стройно изграден от камък мост с единайсет широки свода .
Mazedonisch	На тоа место каде што Дрина изблинува со целата тежина на својата водна маса , зелена и запенета , од привидно затворениот склоп на црните и стрмни планини , стои еден голем , складно изделкан мост од камен , со единаесет широки арки .

Создание письменного Гралис-Корпуса преследует цель (а) собрать и обработать материал для СлЯз, необходимый для анализа отношений

[24] В качестве дополнительного средства для получения и изучения материала можно использовать онлайн-программу по созданию, проведению и обработке опросов под названием „Гралис-Анкетариум" (разработанный в рамках FWF-проекта P19158-G03), а также программу „Гралис-Прескриптариум" по изучению орфографических норм исследуемых языков.

между ними, их совпадений, сходств и различий, а также расстояния. Это подразумевает проведение следующих корпусных процедур: **1.** количественное и качественное сопоставление категорий параллельных фрагментов и единиц текстов, **2.** анализ статистики пар относительно эквивалентных единиц параллельных текстов каждого из языков по ряду параметров (например, по статистике типов соответствия: слово языка^1 – слово языка^2, слово языка^1 – словосочетание языка^2, словосочетание языка^1 – слово языка^2, слово языка^1 – ø языка^2 и т. п.), **3.** интерпретация, качественный анализ статистических данных, **4.** выработка и согласование параметров для последующей корректной сопоставимости получаемых результатов (например, составление частотных словарей словоформ, лексем, морфологических категорий и т. п. для каждого из текстов, для текстов определенного жанра, функционального стиля и всего корпуса текстов данного языка в целом), **5.** исследование внутренних характеристик текстов каждого из языков (интракорреляции) и сопоставление корпусов по ним, а также по типам относительно эквивалентных единиц текстов.

Что касается последовательности процедур, материал сначала вносится в „Roh-Korpus“ („Сырой корпус“) на одном из самых больших серверов Университета Граца (Gedra). Тексты, находящиеся в интернете и не нуждающиеся в решении вопроса авторских прав, включаются в „Warte-Korpus“ („Корпус в ожидании“), расчлененный на пять частей для каждого функционального стиля („FS-Korpus“): литературно-художественный, публицистический, научный, официально-деловой и разговорный. В рубрике „Мета-Корпус“ („Meta-Korpus“) рассматриваются теоретические и практические вопросы создания корпуса и проводится дискуссия между сотрудниками проекта.

Обработка материала проходит в два этапа. На первом отдельно готовятся тексты для каждого славянского языка. Важнейшей частью такой работы является аннотация (паспортизация) – метаязыковая и грамматическая. Метаязыковая аннотация состоит из указания сведений об источнике (авторе, заглавии текста, месте и годе издания, количестве страниц, издательстве, переводе и др.). Грамматическая аннотация указывает на морфологическую структуру, словоизменение и тип сочетаемости с другими словами в рамках словосочетания и предложения. Существуют различные методы и кодировки, используемые для грамматического аннотирования. Так как Гралис-Корпус требует полностью унифицированной аннотации (чтобы искать информацию для всех СлЯз), необходимо выбрать ту, которая больше всего годится для изучаемых языков. Опыт и проведенная работа в процессе подготовки параллельного корпуса для Bs, Hr Sr (Gralis BKS-Korpusa) свидетельствуют о том, что в этих целях самым целесообразным является использование кодировки Multext-East (Multilingual

Texts and Corpora for Eastern and Central European Languages – multilingual dataset for language engineering research and development: MultiText East-www), разработанной в 2004 году группой авторов во главе с Томажем Эрявцем. Так как в рамках Multext-East существует унифицированная система грамматического аннотирования для ряда славянских языков (Bg, Cs, Hr, Ru, Sl, Sr), она будет использована и приспособлена для всех других СлЯз. После проведения разметки текст расчленяется на предложения, в результате чего получится система, в которой каждому предложению языка **A** соответствует предложение языка **B**, **C**... Параллелизация состоит в том, что тексты двух или более СлЯз объединяются, потом проводится их выравнивание. Если, например, в одном языке абзац состоит из трех предложений, а в другом из пяти, приходится устранить такое неравновесие. Для автоматизации данного процесса используются имеющиеся (не)модифицированные разработки. Если существующие программы не в состоянии выполнить поставленные задачи, создаются новые инструментальные средства, позволяющие, в частности, в одном комплексе объединять тексты различных СлЯз, автоматически находить несовпадения в числе предложений и, насколько это возможно, автоматически делать исправления. На этом заканчивается параллелизация и начинается серверная работа, для чего используются IMS Coprus Workbench (CQP) и Asset-Management. Этот этап состоит в том, что из корпусного материала создаются списки языковых единиц, которые затем превращаются при помощи программы MySql в реляционные базы данных, на основании которых можно готовить словари различного типа. IMS Open Corpus Workbench представляет собой набор средств для администрации, подготовки и осуществления поиска в больших текстовых корпусах с лингвистической аннотацией. Его главным компонентом является гибкая и продуктивная поисковая программа CQP (Corpus Query Processor). Первоначально разработанный в Институте машинной обработки языка Штутгартского университета, он был в 2007 году выпущен как программа с открытым кодом (open-source software) с GPL лицензией (GNU General Public License) и размещен на SourceForge. CWB использует для хранения корпуса свой собственный формат. Быстрый доступ достигается за счет бинарной кодировки, полный индекс способствует эффективному поиску словоформ и аннотаций; используются специальные алгоритмы сжатия. В зависимости от аннотации размер корпуса может достигать 500 миллионов словоупотреблений (tokens). CWB содержит инструменты для кодировки, индексации, сжатия, декодирования и частотных распределений, общий реестр, в котором хранится информация о корпусе (название, атрибуты, место нахождения), поисковую программу (CQP), осуществляющую быстрый поиск с использованием синтаксиса регулярных выражений по

значениям атрибутов отдельных позиций (например, по морфологическим тэгам). Поисковая система предоставляет простой и расширенный поиск, который базируется на CQP-синтаксисе (Suche mit CQP-Syntax) и предлагает очень широкие и разнообразные комбинации. Результаты поиска выводятся на монитор вертикально. Название текста/источника обозначается желтым цветом. Нажатием стрелочки с левой стороны заглавия можно получить основную метаинформацию (об авторе, месте и времени издания, издательстве, числе страниц и т. п.). При отображении результатов поиска пользователь может определить размер отображения. Допускаются различные виды сортировки строк, подсчитываются частоты (например, для комбинации слов), составляется многоязычный индекс для параллельных корпусов.

Корпусная работа подразумевает: **1**. выработку инструментальных средств для создания корпуса, его ведения (пополнения, предобработки текстов, паспортизации, контроля их параметров и т. п.), структурирования (разметки структурными пометами), категоризации (качественной квалификации его фрагментов и единиц, выделенных в ходе структурирования текстов), выравнивания и фиксации связи между относительно эквивалентными элементами параллельных текстов корпуса, **2.** создание, тестирование, наполнение, пробную эксплуатацию и развитие и совершенствование оболочки (программного средства) для выполнения работ по созданию, ведению, структурированию, категоризации, выравниванию элементов параллельных текстов, **3.** выработку и согласование схемы метаязыка, а также структурной и категориальной разметки фрагментов и единиц текстов, **4.** переработку текстов каждого из языков, контроль их параметров и характеристик паспортизации, структурную разметку текстов, категориальную разметку фрагментов и единиц текстов (как минимум планируется осуществить лемматизацию и морфологическая квалификацию словоформ), выравнивание (Alignment) соответствующих эквивалентных фрагментов параллельных текстов, **5.** выработку и согласование функционально-стилистической (жанровой) схемы сбора параллельных текстов для корпуса, **6**. сбор и исходную паспортизацию (метаразметку) параллельных текстов по каждому из языков.

10. В процессе создания корпуса кроме подготовки базовых текстов и аннотации метаданных в XML-формате необходимо переработать цифровые ресурсы (MP3-, WAV-данные и т. п.) в рамках Speech-Korpus'a. Asset Management System (AMS) дает такую возможность в форме Workflows.[25] В узком смысле этим понятием обозначаются системы хранения, управления и подготовки цифровых ресурсов, накапливаемых в большом

[25] Развитие устного корпуса (Speech-Korpus) требует больших усилий, много времени и значительных финансовых средств.

количестве. В отличие от Content Management System, в Asset Management выступает на передний план, прежде всего, архивирование постоянных метаданных, доступных для цитирования и гибкого использования цифровых ресурсов. Основная идея Asset Managements состоит в том, чтобы раздробленные цифровые ресурсы получили одну центральную IT-структуру и этим обеспечили продолжительное архивирование имеющегося цифрового фонда знаний, доступного для цитирования. С (текстуально-)технологической точки зрения Asset Management System предоставляет, в первую очередь, очень гибкий способ хранения документов, основанных на XML, и управление корпусом. Именно из-за этих свойств AM-System являются надежным местом для хранения корпусных текстов.

В качестве платформы для трансфера материала используется Open Source Projekt „Fedora" (Flexible Extensible Digital Object Repository Architecture), разработанный в Cornell University (который с 1997 г. вместе финансировали Виргинский университет и Mellon Foundation, а ранее – National Science Foundation). „Fedora" предоставляет структуру для сохранения цифровых web-ресурсов и управления ими (Repository). Она основана на архитектуре SOA. Речь идет о распределяющейся системной архитектуре, не зависящей от платформы и основанной на web-сервисе (SOAP, Simple Object Access Protocol). Она содержит полный текстовый регистр (VolltextIndex), базируемый на Apache Lucene, и версионное управление (Versionsverwaltung) Asset-содержанием (Assetinhalte). Архитектура SOA обладает регистром метаданных с языком запросов ITQL типа SQL (Tucana Technologies), основанном на RDF, и четкой адресацией цифровых ресурсов на базе URL. Для нее характерны тонко расстраиваемые права доступа к Assets и к его частям на основании XACML (Extensible Access Control Markup Language), а также стандартные форматы импорта и экспорта: METS (Metadata Encoding and Transmission Standard). Существует поддержка стандартных протоколов обмена метаданными типа OAI-PMH. На основе носящей технологии (Trägertechnologie) Apache Tomcat можно также осуществить при помощи Repository Clustering загрузочную балансировку (Load Balancing) системного окружения с соответствующим числом пользователей (Concurrent Users).

Для успешного и эффективного проведения корпусных работ существует особый on-line Projekt-Management, в рамках которого созданы базы данных „Персоналиум", „Библиотекариум", а также „Слав-Форум". „Персоналиум" содержит все основные данные о сотрудниках проекта. В эту базу данных вносятся тексты для публикаций, резюме, хендауты и презентации для совместных конференций. В „Библиотекариуме" стоит в распоряжении литература, относящаяся к исследуемой теме. Для постоян-

ного контакта и быстрого обсуждения определенных вопросов используется Slawistik-Forum (http://www-gewi.uni-graz.at/gralis/slaw/forum).

11. В работе над Гралис-Корпусом ожидаются следующие результаты: 1) создание параллельного устного и письменного корпусов для всех СлЯз, снабженного метазыковой и грамматической аннотацией (объем корпуса, широта охвата языков и глубина аннотирования будут зависеть от уровня финансирования), охватывающего все функциональные стили и доступного для всех или как можно большого числа пользователей, при полном соблюдении авторских прав, 2) создание реляционных баз данных – „Slaw-Lexikarium" с корпусными словарными единицами, „Slaw-Grammatikarium" с корпусными грамматическими формами, „Slaw-Akzentarium" с ударениями для всех слов в корпусе, „Slaw-Präskriptarium" с правилами для написания корпусных слов и „Slaw-Bibliothekarium", охватывающий важнейшую литературу.

Литература

Ammon, U. 1986. Explikation der Begriffe 'Standardvarietät' und 'Standardsprache' auf normtheoretische Grundlage. In: Holtus G., Radtke E. (Hg.) *Sprachlicher Substandard.* Tübingen: Max Niemeyer Verlag, 1–63.

Ammon, Ulrich. 1987. Funktionale Typen/Statustypen von Sprachsystemen. In: Ammon U., Dittmar N., Mattheier K.J. (Hg.) *Sociolinguistics/Soziolinguistik.* First Volume / Erster Halbband. Berlin, New York: Walter de Gruyter, 230–263.

Hudson, R. A. 1990 [1980]. *Sociolinguistics.* Cambridge, New York, Portehester, Melbourne, Sydney: Cambridge University Press.

Tošović Branko 2001: *Korelaciona sintaksa.* Projektional. Graz: Institut für Slawistik der Karl-Franzens-Universität.

Tošović B. 2002. *Funkcionalni stilovi. Funktionale Stile.* Graz: Institut für Slawistik der Universität Graz.

Tošović, B. 2008. *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen.* Wien, Berlin: LIT.

Tošović, B. (Hg). 2008a. Das Gralis-Korpus. In: *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen.* Münster: LIT, 724–827.

Tošović, B. 2008b. Das Gralis-Akzentarium. In: *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen.* Münster: LIT, 770–776.

Tošović, B. 2008c. Das Gralis-Bibliothekarium. In: *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen.* Wien, Berlin: LIT, 807–812.

Tošović, B. 2008d. Das Gralis-Präskriptarium. In: *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen.* Wien, Berlin: LIT, 822–825.

Tošović, B. 2008e. Der Unterschied. In: *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen.* Wien, Berlin: LIT, 143–185.

Тошович Б. 2008f. Сопоставительное изучение славянских языков при помощи многоязычного „Гралис-Корпуса". In: *Izučavanje slovenskih jezika, književnosti i kultura kao inoslovenskih i stranih.* Beograd: Slavističko društvo Srbije, 336–340.

Электронные корпусы

Chemnitz German-EnglischTranslation Corpus: http://www.tu-chemnitz.de/phil/InternetGrammar.
Compara-www: http://www.linguateca.pt/COMPARA/index.php.
EPC-www: http://www.statmt.org/europarl.
Gralis-Korpus-www: http://www-gewi.kfunigraz.ac.at/gralis/0.Projektarium/Gralis-Korpus/korpus.html.
Infostream-www: http://ling.infostream.ua/.
Kacenka: http://www.phil.muni.cz/angl/kacenka/kachna.html.
Kollokation-www: http://www.kokken.go.jp/public/world/mirror/www.ids-mannheim.de/gra/kollokation.html.
Leeds-www: http://corpus.leeds.ac.uk.
Lilabar-www: http://lilabar.com/index.php.
Maastr-www: http://www.philhist.uni-augsburg.de/lehrstuehle/anglistik/.
Opus-www: http://opus.lingfil.uu.se.
RNK-www: http://corpora.yandex.ru/search-para.html.
RPC-www:http://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/rund-ums-institut/regensburger-korpora/index.html.
Ruscorpora-www: http://ruscorpora.ru; http://ruscorpora.ru/search-para.html.
SPI-www: http://nevmenandr.net/slovo.
Traumdeutung-www: http://www.aac.ac.at/lab_parallel_freud.html.