МАПРЯЛ

КАФЕДРА СЛАВИСТИКИ БЕЛГРАДСКОГО УНИВЕРСИТЕТА
СЛАВИСТИЧЕСКОЕ ОБЩЕСТВО СЕРБИИ

МЕЖДУНАРОДНЫЙ СИМПОЗИУМ
посвященный
130-летию Кафедры русского языка и литературы
и
60-летию Славистического общества Сербии
(Белград, 3–5 июня 2008)

ИЗУЧЕНИЕ СЛАВЯНСКИХ ЯЗЫКОВ, ЛИТЕРАТУР И КУЛЬТУР КАК ИНОСЛАВЯНСКИХ И ИНОСТРАННЫХ

Сборник докладов

Редактор
БОГОЛЮБ СТАНКОВИЧ

Помощники редактора
Вера Борисенко-Свинарская,
Петар Буняк, Вукосава Джапа-Иветич

СЛАВИСТИЧЕСКОЕ ОБЩЕСТВО СЕРБИИ
БЕЛГРАД
2008

МАПРЯЛ

КАТЕДРА ЗА СЛАВИСТИКУ УНИВЕРЗИТЕТА У БЕОГРАДУ
СЛАВИСТИЧКО ДРУШТВО СРБИЈЕ

МЕЂУНАРОДНИ СИМПОЗИЈУМ
посвећен
130-годишњици Катедре за руски језик и књижевност
и
60-годишњици Славистичког друштва Србије
(Београд, 3–5 јун 2008)

ИЗУЧАВАЊЕ СЛОВЕНСКИХ ЈЕЗИКА, КЊИЖЕВНОСТИ И КУЛТУРА КАО ИНОСЛОВЕНСКИХ И СТРАНИХ

Зборник реферата

Приређивач
БОГОЉУБ СТАНКОВИЋ

Помоћници приређивача
Вера Борисенко-Свинарски,
Петар Буњак, Вукосава Ђапа-Иветић

СЛАВИСТИЧКО ДРУШТВО СРБИЈЕ
БЕОГРАД
2008.

Б. Тошович (Австрия)

СОПОСТАВИТЕЛЬНОЕ ИЗУЧЕНИЕ СЛАВЯНСКИХ ЯЗЫКОВ ПРИ ПОМОЩИ МНОГОЯЗЫЧНОГО „ГРАЛИС-КОРПУСА“

0. Для изучения славянских языков используются в настоящие время самые современные средства. Одними из них являются электронные корпуса, при помощи которых можно в большом количестве текстов искать и находить необходимую информацию об употреблении слов, форм и конструкций.

Существуют два основных типа электронных корпусов. Первый из них – это одноязычные корпуса, предназначенные для изучения только одного языка. Такими для русского языка являются „Национальный корпус русского языка“ и „Национальный корпус русского литературного языка (Наруско)“, а для сербского „Корпус савременог српског језика на Математичком факултету Универзитета у Београду“. Другой тип – двуязычные и многоязычные электронные корпуса, предназначенные для сопоставительного изучения языков. Они обычно называются параллельными корпусами.

Создание электронных словарей является очень трудоемкой работой. Сложность их создания удваивается при подготовке параллельных корпусов, так как необходимо связать в одно целое два, три, четыре языка, а то и несколько языков. По этой причине таких корпусов очень мало.

Одним из параллельных корпусов является „Гралис-Корпус“, который создан в 2007 году. Для него в качестве платформы используется лингвистический славистический портал Университета Грац „Гралис“ (http://www-gewi.kfunigraz.ac.at/gralis/). „Гралис-Корпус“ состоит из нескольких корпусов, охватывающих как минимум два языка. Первый такой корпус развит для изучения сербского, хорватского и бошняцкого языков. Он состоит из двух частей – письменного корпуса (Text-Korpus) и устного корпуса (Speech-Korpus).

Второй корпус, над которым мы работаем, включает два близкородственных южнославянских языка – болгарский и македонский. Третий предназначен для изучения восточнославянских языков – русского и украинского.

Как видно, при создании многоязычного славянского корпуса „Гралис“ упор делается на близкие славянские языки, особенно на те, которые своими взаимоотношениями вызывают особый лингвистический и общественный интерес (самыми типичными примерами являются, с одной стороны, сербский, хорватский и бошняцкий, и, с другой, болгарский и македонский).

1. Славяно-славянских параллельных корпусов очень мало. Здесь можно упомянуть „Русско-словацкий параллельный корпус“, который в настоящее время содержит небольшое количество слово: в словацкой части 818 097 словоупотреблений, 43 381 предложений, а в русской части 819 09 слов и 46 832 предложений.

Среди неславянских параллельных корпусов выделяется Europarl Parallel Corpus: http://www.statmt.org/europarl/, в рамках которого развиты подкорпуса Danish-English, German-English, Greek-English, Spanish-English, Finnish-English, French-English, Italian-English, Dutch-English, Portuguese-English, Swedish-English. Центр теории перевода Университета Leeds развил „Leeds Corpus“ (http://corpus.leeds.ac.uk/), охватывающий различные языки (английский, китайский, французский, немецкий, итальянский, японский, испанский), русский, который 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). Английский корпус состоит из новостей агентства Reuters. Использование корпуса ограничено только в целях исследования. Следующим параллельным корпусом является „MAASTR“ (http://www.philhist.uni-augsburg.de/lehrstuehle/anglistik/sprachwissenschaft/mitarbeiter/stoll/elekhilf/), охватывающий мастристские соглашения на немецком и английском языках. Некоторые из этих корпусов являются недоступными из-за проблем с авторскими правами. Создан также параллельный корпус в рамках немецко-французского проекта „Коллокации в контексте“. Этот корпус охватывает немецкие тексты с французским переводом и французские тексты с немецким переводом. Его объем – 15 миллионов словоупотреблений. Он состоит из CELEX-документов (право Европейского Содружества – соглашения, внешние отношения, законы) и документов Европейского парламента (EUROPARL). Ожидается его развитие до 50 миллионов словоупотреблений для каждого языка. На основе этого корпуса сделан „Немецко-французский словарь коллокаций“ – словарь сочетаемостей, охватывающих типичное и постоянное окружение какого-либо лексического элемента, в первую очередь прилагательных и существительных (http://www.kokken.go.jp/public/world/mirror/www.ids-mannheim.de/gra/kollokation.html). К этой группе относится открытый online корпус под названием „Parallel Corpus of Portuguese and English“, сокр. COMPARA (http://adamastor.linguateca.pt/COMPARA/construcao_compara.php). В 1999 году началась в Институте немецкого языка в Мангейме (Institut für Deutsche Sprache, IDS, Mannheim) работа над проектом „GeFrePac (German-French Reciprocal Parallel Corpus), финансируемым ELRA (European Language Resources Agency, Paris) и IDS, под руководством Вольфганга Тойберга (Wolfgang Teubert).

2. Целью настоящего проекта является создание параллельного корпуса для изучения и преподавания славянских языков на интракоррелационном уровне (отношения внутри каждого языка в отдельности), интракорреляционном уровне (отношения между особо близкими славянскими языками), экстракорреляционном уровне (отношения между любыми славянскими языками) и супракорреляционном уровне (отношения между славянскими и неславянскими языками). Славянский Гралис-Корпус предназначен для тех, кто занимается (1) славянскими языками в целях обучения и исследования, (2) отношениями между славянскими литературами и культурами, (3), славянскими странами.

3. Многоязычный славянский Гралис Корпус будет охватывать все функциональные стили (литературно-художественный, научный, публицистический, официально-деловой и разговорный) и будет сбалансированным (будет отражать общую структуру двух языков).

4. В качестве платформы для представления и использования корпусного материала, его подготовки и обработки, а также для коммуникации участников в проекте, будет использован лингвистический славистический портал Университета Грац „Gralis“ (http://www-gewi.kfunigraz.ac.at/gralis/).

5. Для создания параллельного славянского „Гралис ДеРу-Корпуса“ существуют необходимые предпосылки. В Институте славистики и Институте переработки информации Университета Грац уже накоплен довольно хороший опыт и знания по составлению параллельных корпусов. В рамках трехлетнего проекта „Различия между боснийским/бошняцким, хорватским и сербским языками“ („Die Unterschiede zwischen dem Bosnischen/Bosniakischen/Kroatischen und Serbischen“ (FWF-Projekt, P19158-G03, 2006-2009), который финансирует Австрийский Фонд научных исследований (FWF), развит усилиями сотрудников Института славистики и Института обработки информации электронный корпус параллельных текстов для изучения боснийского/бошняцкого, хорватского и сербского языков под названием „Gralis BKS-Korpus“. Он состоит из Текстового корпуса (Text-Korpus) и Устного корпуса (Speech-Korpus). Текстовый корпус насчитывает более трех миллионов словоформ (tokens). Устный корпус охватывает около 300 записей устной речи (столько же находится в подготовке). Оба подкорпуса хорошо функционирует и используются для исследования и обучения (они теоретически рассмотрены и описаны в ряде публикаций). Так как славянскийй Гралис-Корпус задуман как многоязычный, полифункциональный, многоуровневый и гибкий корпус параллельных корпусов, не зависящий от типа языка и его графической системы, его можно расширять и включать самые различные языки. Одной из целей является создание на базе данного корпуса системы различных электронных словарей слаянско-славянских. Кроме того, будет создана основа для нового подхода к машинному переводу (не на базе лексического материала, а на основе большого и разнообразного текстового массива) и распознаванию речи. Эта ориентация укладывается в новые поиски, в том числе в так. наз. Translation memory, суть чего состоит в том, что в компьютерной памяти накапливаются переводы предложений и из нее извлекаются эквиваленты.

6. Работа над созданием даннгого корпуса охватывает два основных процесса – накопление материала и его обработку. Это направление преследует следующие цели: **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 структуировнаия текстов), выравнивания и фиксации связи между относительно эквивалентными элементами параллельных текстов корпуса в ходу полуавтоматических процедур, перемежаемых ручным контролем и экспертными решениями алгоритически неясных соотношений фрагментов текстов, **2.** создание, тестирование, наполнение, пробную эксплуатацию и развитие оболочки (программного средства) для выполнения работ по созданию, ведению, структурированию, категоризации, выравниванию элементов параллельных текстов, **3.** выработку и согласование принципиальной схемы метаязыка, а также структурной и категориальной разметки фрагментов и единиц текстов, **4.** переработку текстов каждого из языков, контроль их параметров и характеристик паспортизации, полную структурную разметку текстов, категориальную разметку фрагментов и единиц текстов (как минимум предусматривается осуществить лемматизацию и морфологическая квалификацию словоформ), выравнивание (Alignment) соответствующих эквивалентных фрагментов параллельных текстов, **5.** выработку и согласование принципиальной жанровой схемы сбора параллельных текстов для корпуса, 6. сбор и исходная паспортизацию (метаразметку) параллельных текстов по каждому из языков.

7. Так как в Гралис-Корпус будут вноситься только тексты, на использование которых имеется согласие носителей авторских прав, решение этого вопроса будет первой предпосылкой для их включения. Тексты, на которые получено согласие, сначала будут вноситься в так. наз. Roh-Korpus („Сырой корпус“). Для него уже выделено специальное место на одном из самых больших серверов Университета Грац (Gedra). Тексты, находящиеся в интернете и не нуждающиеся в решении вопроса авторских прав (так как они уже опубликованы), будут вноситься в „Warte-Korpus“ („Корпус в ожидании“), в его часть „FS-Korpus“ (с указанием на линк), расчлененный на пять функциональных стилей (оттуда и название „FS-Korpus“) – литературно-художественный, публицистический, научный, официально-деловой и разговорный. В рубрике „Мета-Корпус“ („Meta-Korpus“) будут рассматриваться теоретические и практические вопросы создания „Гралис-Корпуса“ и будет проводиться дискуссия между сотрудниками на проекте.

8. Обработка материала будет проходить в двух этапах. На первом будут отдельно готовиться тексты на славянских языках. Важнейшей частью такой работы является аннотация – метаязыковая, лексико-семантическая, грамматическая и стилистическая. Метаязыковая (паспортизация) состоит из указания об источнике (авторе, заглавии текста, месте и годе издания, количестве страниц, издательстве, переводе и др.). Лексико-семантическая аннотация указывает на основные лексические и семантические характеристики слова (напр. разг., устар., новое и т. п.). Грамматическая аннотация показывает, какая морфологическая структура и тип сочетаемости с другими словами на уровне словосочетания, предложения и текста. Стилистическая аннотация указывает на функциональный стиль, подстиль и жанр. После проведения аннотации текст будет расчленяться на предложения, в результате чего получится система, в которой каждое предложение будет находиться в отдельной ячейке. Этим заканчивается обработка текста и начинается параллелизация. Она состоит в том, что тексты на различных языках объединяются, потом проводится их выравнивание, для того чтобы получилась схема: предложение языка **A** – предложения языка **B**. Если, например, в одном языке абзац состоит из трех предложений, а в другом из пяти, приходится устранить такое неравновесие. Для автоматизации данного процесса. будут использованы уже имеющиеся (не)модифицированные разработки Если существующие программы не в состоянии выполнить поставленные задачи, будут созданы новые инструментальные средства, позволяющие в частности в одном комплексе объединять разноязычные тексты, автоматически находить несовпадения в числе предложений и, насколько это возможно, автоматические делать исправления. На основе трех программах, которые используются в настоящее время для параллеллизации боснийских/бошняцких и хорватских текстов – Аннотаторе (Annotator), Верификаторе (Verifikator) и CkeckScript-е – будет сделана попытка объединить все три (если не появится у других разработки, эффективно и надежно решающие данные вопросы). В эту программу было бы целесообразно объединить и процедуры по обработке XML-файлов и созданию файлов TEI-формата. Этим заканчивается параллелизация и начинается серверная работа, для чего будут использованы IMS Coprus Workbench (CQP) и Asset-Menagement.

9. IMS Open Corpus Workbench представляет собой набор ср...ств 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 [Christ 1994], [Christ, Schulze 1995], в 2007 году он был 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 слов. CWB содержит следующие компоненты: инструменты для кодировки, индексации, сжатия, декодирования и частотных распределений, общий реестр, где хранится информация о корпусе (название, атрибуты, место нахождения), поисковую программу (CQP), которая осуществляет быстрый поиск с использованием синтаксиса регулярных выражений по значениям атрибутов отдельных позиций (например, по морфологическим тэгам). Система поиска информации будет предоставлять две возможности: простой и расширенный поиск. Для простого поиска (einfache Suche) используются различные комбинация типа „.*“, а расширенный будет базироваться на CQP-синтаксисе (Suche mit CQP-Syntax) и будет предлагать очень широкие и разнообразные комбинации. При этом можно будет выбирать первичный язык. Результаты поиска будут вертикально выводиться на монитор. Название текста/источника будет обозначено желтый цветом. Нажатием стрелочки с левой стороны заглавия можно будет получить основную мета-информацию (об авторе, месте и времени издания, издательстве, числе страниц и т. п.). При отображении результатов поиска пользователь может определить размер отображения „ключевое слово в контексте“. При этом допускаются различные виды сортировки строк „ключевое слово в контексте“, подсчитываются частоты (например, для комбинации слов), составляется многоязычный индекс для параллельных корпусов.

ЛИТЕРАТУРА

1. Tošović 2008e: **Tošović, Branko.** Das Gralis-Korpus // *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen* – Wien: LIT, 2008. – S. 724–749.

СРПСКА КЊИЖЕВНОСТ НА СТРАНИЦАМА ПОЉСКОГ ЧАСОПИСА „KAMENA” (1933–1993)

Пољска периодика постала је извор знања о српској књижевности већ тридесетих година XIX века[1]. Од тада, пре свега, на пољској територији под аустријском окупацијом, где се тенденција зближавања словенских народа схватала као нека врста отпора против германизације[2], почели су се појављивати преводи српских народних песама и први чланци о књижевности.

Завршетак Првог светског рата и настанак обновљених, самосталних држава – Републике Пољске (Rzeczpospolita Polska) и Краљевине СХС[3] проузроковао је развој пољско–југословенских веза у многим областима и допринео ширењу код Пољака знања о историји и култури новонастале јужнословенске државе.

У укључивању Пољака у проблематику српске (и осталих југословенских књижевности) битну улогу одиграла је меродавна штампа која је већ дуже времена постојала на издавачкој сцени[4], додаци уз популарне листове[5] и новонастали часописи[6]. Међу њима од великог значаја била је *stricte* славистичка штампа (нпр. „Kultura Słowiańska”, „Przegląd Polsko-Jugosłowiański”, „Ruch Słowiański”, „Panslavia”) и периодика која се упркос „варшавоцентризму” развијала у пољској провинцији[7].

Овој групи малотиражних часописа припада и „Kamena“ (1933–1993). Одлуку о издавању у Хелму (Chełm), градићу на истоку Пољске овог месечника, донели су, упркос материјалним тешкоћама, два учитеља тамошње гимназије – песник Казимјеж Анџеј Јаворски (1897–1973) и цртач, песник и преводилац Зенон Вашњевски (1891–1945).

„Kamena”, чију је проблематику наговештавао већ наслов[8] часописа, била је од свог настанка у јесен 1933. године конципирана као форум за популаризацију савременог песништва и везана углавном за пољске авангардне и левичарске писце, а њени оснивачи истицали су свој умерени авангардизам и антифашистичке погледе. Кругу сарадника током постојања часописа припадали су: Јузеф Чехович (1903–1939), Бру-

[1] М. Јакубјец-Семковова, *Шта су Пољаци пре сто година знали о српској књижевности*, у *Сто година полонистике у Србији*, Београд 1996.

[2] Z. Niedziela, *Słowiańskie zainteresowania pisarzy lwowskich w latach 1830–1848*, Kraków 1966, стр. 12.

[3] П. Буњак, *Преглед пољско-српских књижевних веза (до II светског рата)*, Београд 1999, стр. 75.

[4] Овим листовима припадали су: „Tygodnik Ilustrowany“ који је почео излазити 1859., „Bluszcz“ штампан од 1865., „Świat” из 1906.); упор. A. Zawada, *Dwudziestolecie literackie*, Wrocław 1995, стр. 148.

[5] Најважнији у овом погледу били су: „Kurier Literacko-Naukowy“ – додатак уз „Ilustrowany Kurier Codzienny“ и „Żagary” (1931-1932) – додатак уз „Słowo“ који се претворио у самостални књижевно-уметнички месечник]; упор. A.Zawada, *Dwudziestolecie literackie*, Wrocław 1995, стр. 150–151.

[6] Најважнији у том погледу био је културно-књижевни недељник „Wiadomości Literackie“ (1924–1939), важну улогу на књижевној сцени одиграли су „Skamander“ (1920–1928 и 1935–1939), „Zwrotnica“ (1922–1923 и 1926–1927), „Linia“ (1931–1933), „Kwadryga“ (1927- 1931), „Dźwignia“ (1927–1928), „Pion“ (1933–1939), „Prosto z mostu“ (1935) итд.

[7] Један од цењених часописа „Okolica Poetów" песник Станислав Черник (1899-1969) уређивао у Остжешову; Упор. A. Zawada, *Dwudziestolecie literackie*, стр. 174.

[8] У римској митологији „Camenae” означавала је грчку Музу која отелотворује поезију.