РЕЦЕНЗИИ И ОБЗОРЫ

МЕЖЪЯЗЫКОВЫЕ КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ И ПРИЛОЖЕНИЯ ДЛЯ СЛАВЯНСКИХ ЯЗЫКОВ: ТЕХНОЛОГИЧЕСКИЕ И ГЕОПОЛИТИЧЕСКИЕ АСПЕКТЫ[1]

Е.Паскалева
София, Болгария

Языки в лингвистике классифицируются по семьям на основании типологических признаков (например, семья славянских языков), а также по союзам (например, балканский языковой союз). Языки в геополитике делятся на основании участия стран, в которых они приняты в качестве официальных, в геополитических структурах. Это деление не политическое, в том смысле, который предполагает квалификацию или элемент соревнования: является ли данная страна полноправным членом, недавно принятым членом, в скором будущем будет принята, никогда не будет принята, есть ли у нее шансы на членство вообще и т.п.

Геополитический аспект языкового деления касается только вопроса включения соответствующего языка в документную базу соответствующей структуры, а также возможности осуществления более интенсивного языкового обмена и более активной переводческой деятельности, и далее – возможности разработки более масштабных многоязычных приложений в одной из самых модных областей современных коммуникационных технологий – области языковых технологий (language technologies).

История и предпосылки

Языковые технологии (в их мультиаспекте) развиваются особенно успешно в процессе совместной научной и технологической деятельности, которая осуществляется по линии совместных инициатив – общих проектов и производств.

За последние 15 лет в такой деятельности Болгария участвовала, как правило, по программам Европейской комиссии – сначала как член группы *Восточные страны*, потом в качестве ассоциированного члена, а в течение последнего года и в качестве полноправного члена общности.

[1] Оригинальный болгарский текст настоящей публикации был прочитан в качестве доклада на Международной конференции «Кириллица в глобализирующемся мире», проведенной 6 ноября 2007 г. в Софии в рамках программы мероприятий Года русского языка.

Для некоторых болгарских научных звеньев, имевших предшествующий опыт и задолго до того заложенные основы (в частности, в области компьютерных технологий), такое участие было возможным и нетрудным для осуществления, особенно если иметь в виду установленные научные связи и совместную деятельность с соответствующими партнерами.

Так обстояли дела и с *Секцией лингвистического моделирования*, созданной в 1987 г. в Институте параллельной обработки информации Болгарской академии наук (ранее это структурное звено называлось *Лабораторией лингвистического моделирования* и функционировало в составе Координационного центра информатики и вычислительной техники при БАН)[2].

Лаборатория поддерживала связь с Российской академией наук (через Компьютерный фонд русского языка при Институте русского языка). «По старой дружбе» нам был предоставлен электронный вариант известного всем русистам Грамматического словаря А.А.Зализняка [Зализняк 1977]. Программисты обоих научных институтов осуществили его загрузку на платформу Windows. Этот словарь, включающий в себя свыше 100 000 единиц, снабженных богатой грамматической информацией, стал основой для многих будущих болгаро-российских разработок.

В период 1995 – 2005 гг. Секция лингвистического моделирования принимала активное участие во многих европейских проектах по программам ТЕМПУС, КОПЕРНИК, 6-ой рамочной программе, а также в отдельных двусторонних проектах. К сожалению, тогда, как и теперь, болгаро-российские совместные инициативы были регламентированы и спонсированы символично только программами БАН по безвалютному эквивалентному обмену. По линии этих программ нельзя получить средства на создание больших компьютерных приложений, требующее немалых финансовых затрат.

С 2005 г. до настоящего момента, в контексте будущего и настоящего членства Болгарии в европейских структурах, исследования в области компьютерной обработки болгарского, а также других славянских языков получают определенное финансирование по линии европейских проектов. Однако в этих проектах (по причинам, которые будут изложены ниже) русский язык, так же как и языки стран, не являющихся членами ЕС, не фигурирует в проектных заданиях.

Стремление к унификации компьютерных стандартов, в частности стандартов для представления языкового знания, позволило нам включить –

[2] Лаборатория (благодаря своим членам-основателям) опиралась на серьезную традицию в области автоматической обработки языка и машинного перевода. Начало этой традиции было положено Группой по проблемам машинного перевода, созданной в 1964 г. при Математическом институте БАН и работающей под руководством Александра Людсканова. Первым экспериментом группы явилась система русско-болгарского машинного перевода, реализованного на машине «Минск-2» (см. [Paskaleva 2002]).

на добровольном основании – в эти стандарты языки, находящиеся за рамками программ, например русский, сербский, македонский и др.

Создание общей парадигмы компьютерной обработки славянских и балканских языков

Болгарский ученый со славистической подготовкой, работающий над созданием современных компьютерных средств обработки родного языка и его кириллицы, всегда обращен в сторону двух основных источников знания и языковых ресурсов (кроме ресурсов на родном языке). Этими источниками являются:

- европейские языковые стандарты и ресурсы, созданные в рамках общих научных инициатив и на основе богатой документной базы административной документации общности;
- русские и славянские языковые ресурсы, языковедческая традиция и общая проблематика, базирующаяся на типологической близости языков.

Эти два источника (источники методов обработки и источники языковых ресурсов), однако, пересекаются в одном сравнительно малом по объему подмножестве. Это связано, прежде всего, с отсутствием протекций и с тем, что еще не создана общая основа исследований и их спонсирования.

Миссия болгарского исследователя, работающего в этом направлении, заключается в том, чтобы попытаться сократить это расстояние, построить своего рода мост между двумя общностями. Они, естественно, не изолированы друг от друга, но связи между ними фрагментарны, спорадичны, так как осуществляются на основе отдельных научных контактов – межуниверситетских, межакадемических и личных. Мост, который необходимо построить между двумя общностями, будет служить, прежде всего, нам самим, нашему профессиональному самоутверждению. Ни одно общее научное мероприятие, связанное со славянскими языками в Европе, не может считаться по-настоящему славистическим, если в него не включен и русский язык.

Не всегда эта миссия болгарского ученого успешна, но затраченные усилия не напрасны – по крайней мере, в отношении собственного научного роста. Области, в которых Секция лингвистического моделирования попыталась заполнить пустое пространство, связаны с:

- общими грамматическими инструментами электронной обработки текстов;
- общей базой больших текстовых корпусов, обработанных для целей межъязыковых исследований и приложений.

Общие грамматические инструменты для автоматической обработки

После переформатирования упомянутого *Грамматического словаря русского языка* в соответствии со стандартами французской системы обработки корпусов INTEX [Silberztein 1993] появилась возможность создавать одновременно для болгарского и русского языков анализирующие грамматические программы, софтвер для подготовки так называемого учебного корпуса для снятия грамматической многозначности, а также

возможность проведения статистических экспериментов с целью измерения лексической близости между двумя языками (в основном для обнаружения так называемых *когнатов*, в том числе и ложных – «ложных друзей переводчика»).

Следующим шагом на пути создания моста, связывающего нас с *другой общностью*, было переформатирование этого словаря в соответствии с международным стандартом для грамматической аннотации, созданным в рамках европейского проекта Multext-East [MTE 2004]. В этом проекте восточное расширение включает болгарский, чешский, сербский, хорватский и словенский языки. Подобная стандартизация с расширением (для переводных текстов) была осуществлена и в проекте INTERA [INTERA 2003].

Это расширение в сторону восточноевропейских и балканских языков становится естественным в связи с начавшимся процессом включения в ЕС некоторых стран и будущими заявками других. Русский компонент в этом стандарте все еще находится за рамками существующей схемы, но мы рады предоставленной нам возможности проведения новых компьютерных типологических исследований, а также нашему добровольному вкладу в эту область и результатам пионерских усилий (то и другое обычно идут рука об руку).

Общая текстовая база и межъязыковые исследования

В современных языковых компьютерных технологиях основным ресурсом для проведения исследований и создания приложений служат параллельные корпуса большого объема. Речь идет об электронных переводных текстах объемом в миллионы слов, которые обрабатываются особым образом – элементы текста размещаются в соответствии с элементами его перевода.

Естественным источником таких корпусов служат большие коллекции переведенных текстов, как правило административных документов соответствующей европейской структуры. Таким большим корпусом документов 22 стран-членов ЕС является, например, коллекция, представленная в [JRC-Acquis 2007], в которой есть и болгарский компонент.

Операция по параллелизации и выравниванию текстов осуществляется с помощью специального софтвера, так называемых программ-выравнителей. Первая такая программа была создана в 1995 г. членами Секции лингвистического моделирования в рамках европейского Коперникус-проекта GLOSSER [Nerbonne et al. 1997].

С помощью этой программы были обработаны сотни тысяч слов русских текстов, переведенных на болгарский язык. Программа работает с кириллицей в стандарте Windows 1251. На входе подаются пары электронных файлов. Таким образом были обработаны тексты «Мастера и Маргариты» М.Булгакова, «Золота партии» И.Бунича, некоторых научно-фантастических произведений и др.

За последние несколько лет в Европейском союзе в связи с накоплением общей европейской документной базы, локализованной в основном в пространстве Интернета, возникла необходимость создания новых программ-выравнителей, которые должны обрабатывать в пакетном режиме

большой объем переводных текстов. Это привело к созданию нового типа программы-выравнителя, работающей в пакетном режиме и извлекающей текстовый материал объемом в тысячи файлов из веб-сайтов. Так как тексты, извлекаемые из веб-сайта, известного славистам под названием Balkan-Times, созданы посредством трех шрифтов – латинского, греческого и кириллического, – программа приспособлена для обработки именно этих шрифтов. В ней предусматривается возможность выбора разных кодировок, например Windows 1251, а также Unicode стандартов для кириллицы [Genov 2007].

Упомянутый сайт – www.setimes.com – поддерживается Министерством обороны США и содержит новости, опубликованные начиная с 2002 г. на девяти южнославянских и балканских языках (болгарском, сербском, македонском, хорватском, албанском, турецком, греческом, румынском, боснийском), а также на английском. Объем текстов для каждого из языков превышает 3 миллиона словоупотреблений.

На материале этого корпуса были проведены эксперименты по выравниванию (на уровне предложения) следующих пар языков: болгарский – английский, болгарский – греческий, болгарский – македонский.

Так как речь идет об одном материале, представленном на разных языках, корпус является бесценным источником для сопоставлений, в частности для работ по сравнительной лексикологии и исследований по теории и практике перевода.

На последней большой международной конференции «Современные методы в автоматической обработке текстов» (Болгария, Боровец, 2007) обсуждались результаты таких исследований, как:

- исследования лексической близости между болгарским и македонским языками (на материале этого корпуса) [Nakov et al. 2007-3];
- исследования по автоматической классификации (clusterization) болгарских и английских текстов (на материале этого корпуса) [Alfred et al. 2007];
- исследования лексической близости между русским и болгарским языками (на другом языковом материале) [Nakov et al. 2007-2, 2007-3].

К сожалению, в этом корпусе отсутствует русский компонент. До 2005 г. тексты переводились и на сербскую кириллицу, в настоящее время замененную латиницей.

Сравнительная таблица имеющихся в наличии общих инструментов – программ и корпусов (для группы языков, объединенных и частично пересекающихся по географическим и языковым признакам) обобщенно представляет ситуацию, объединившую языковые (типологические) и геополитические критерии общности и совместной работы.

язык	ЕС член	Se-times	ACQUIS	INTERA	MTE	Словари
Албанский		x				
Румынский	x	x	x		x	x
Греческий	x	x	x	x	x	x
Турецкий		x				x
Болгарский	x	x	x	x	x	x
Македонский		x			x	
Сербский		x		x	x	x
Хорватский		x			x	x
Боснийский		x				
Словенский	x		x	x	x	x
Чешский	x		x		x	x
Русский					x	x
Венгерский	x		x		x	x
Эстонский	x		x		x	x
Английский	x	x	x	x	x	x

Возможности расширения «миссии»

Пустые участки в таблице, связанные, прежде всего, с отсутствием параллельных корпусов русского языка (не только в паре русский – болгарский, но и в парах с другими славянскими языками), не могут быть заполнены административными текстами геополитической общности, так как такой общности пока не существует.

Одним из возможных способов заполнения пустых участков является составление параллельных корпусов, включающих переводы русского оригинала на несколько славянских языков.

Такая работа уже ведется, пока в виде небольшого эксперимента. Идет, например, поиск переводов «Мастера и Маргариты» Булгакова на другие славянские языки. Были проведены эксперименты не на базе пары оригинал – перевод, а на базе пары перевод – перевод (болгарский и македонский переводы книги) [Pacovski 2006].

Такой тип деятельности предполагает участие добровольцев из Болгарии и других стран.

Проблема состоит в том, что степень доступности электронных файлов – оригиналов и переводов – различается. Составление пар, троек, четверок (оригинала и его переводов) вопреки логике необходимо начинать не с электронного файла оригинала. Если взять за основу русский язык и искать переводы с русского на другие славянские языки, самым доступным окажется оригинал (поскольку в Рунете существуют большие электронные библиотеки с многочисленными ресурсами на русском языке). Сложнее будет найти переводы, так как электронные библиотеки содержат, как правило, меньше переведенных произведений, да и строгость копирайта увеличивается в западном направлении.

Как добровольцы могли бы быть использованы студенты-слависты.

Особую надежду мы связываем со студентами, которые посещали в этом году новый элективный курс, проведенный в рамках бакалавриата на

факультете славянских филологий Софийского университета. К сожалению, в виду занятости одного из лекторов – доц. Т.Августиновой (преподавателя кафедры славистики университета в Саарбрюкене) курс был интенсивным и продолжался всего две недели. Надеемся, что в следующем семестре слушателей этого курса будет еще больше.

В заключение еще раз подчеркнем, что проводить славистические типологические исследования в очерченной геополитической ситуации могут только исследовательские группы, созданные на добровольной основе в ожидании подходящего формата для совместной деятельности и помощи со стороны научных институтов или государства.

ЛИТЕРАТУРА

Зализняк 1977 – *Зализняк А.А.* Грамматический словарь русского языка. 4-е изд. М., 1977.

Alfred et al. 2007 – *Alfred R., Paskaleva E., Kazakov D., Bartlett M.* Hierarchical, agglomerative clustering for cross-language Information Retrieval // International journal of translation. Vol XX, 2007.

Genov 2007 – *Genov N.* LORA – a basic tool for creation and primary processing of multilingual Balkan text corpora aligned to English. Presentation on BIS21++ Information days – May 2007, Hissar.

INTERA 2003 – http://www.elda.org/intera

JRC-Acquis 2007 – http://langtech.jrc.it/JRC-Acquis.html

MTE 2004 – MULTEXT-East Morphosyntactic Specifications. Version 3.0, May 2004 – http://nl.ijs.si/ME/V3/msd/msd2.pdf

Nakov et al. 2007-1 – *Nakov P., Pacovski V., Paskaleva E.* Extracting Translation Lexicons from Bilingual Corpora: Application to South Slavonic Languages // A common natural language processing paradigm for Balkan languages. Workshop, RANLP-2007, September, 2007.

Nakov et al. 2007-2 – *Nakov S., Nakov P., Paskaleva E.* Cognate or False Friend? Ask the Web! // A Workshop on Acquisition and Management of Multilingual Lexicons, RANLP 2007, September 2007.

Nakov et al. 2007-3 – *Nakov P., Nakov S., Paskaleva E.* Improved Word Alignments Using the Web as a Corpus // Recent advances in Natural Language Processing, RANLP 2007, September 2007.

Nerbonne et al. 1997 – *Nerbonne J., Karttunen L., Paskaleva E., Proczeky G., Roosmaa T.* Reading more in Foreign Languages // Fifth Applied Natural Language Processing Conference, April 1997. Washington, ACL – http://ucrel.lancs.ac.uk/acl/A/A97/A97-1020.pdf

Pacovski 2006 – *Pacovski V., Paskaleva E.* Aligning the Translations – a Possible Strategy for Creation of Aligned Corpora (for South-Slavic languages) // Formal Approaches to South Slavic and Balkan Languages (FASSBL-5). Sofia, 2006.

Paskaleva 2002 – *Paskaleva E.* Alexander Ljudskanov // Early years in machine translation. Memories and bibliographies of pioneers. ed. W.J.Hutchins, John Benjamins, 2000. Amsterdam/ Philadelphia, pp. 361-377. Болгарский

перевод: Езикови технологии и софтуерен пазар. Euromap project. София, 2003.

Paskaleva 2007 – *Paskaleva E.* Balkan – South East Corpora Aligned to English // A common natural language processing paradigm for Balkan languages. Workshop, RANLP-2007, September, 2007.

Paskaleva et al.1997 – *Paskaleva E., Michov St.* Second Language Acquisition from Aligned Corpora // Proc. of the Int. Conf. Language Technology and Language Teaching. Groningen, The Netherlands, April 1997.

Paskaleva-Avgustinova 2007 – *Паскалева Е.* (ИПОИ–БАН), *Августинова Т.* (Saarland University). Компютърна компаративистика (езикови технологии за славянските езици) – http://www.coli.uni-saarland.de/%7etania/block-kurs/

Silberztein 1993 – *Silberztein M.* Dictionnaires électroniques et analyse automatique de textes: le système INTEX. Masson Ed. Paris, 1993.

Перевод с болгарского А.Липовской