Мария Шимкова
(Братислава)

ЯЗЫК – КОРПУС – СЛОВАРЬ

Abstract: This article provides a brief overview of the Slovak dictionaries, lexical databases and corpora produced at the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences, especially in the department of the Slovak National Corpus. The paper deals with the nature and importance of language corpora and their use. It next explains the need for close cooperation between corpora linguistics and (computational) lexicography. At present, the links between natural language processing and lexicography are in Slovakia well developed and mutually beneficial.

Key words: language, corpora, dictionary, computational lexicography

В честь мероприятия, посвящённого 70-летнему юбилею Института болгарского языка и болгарской академической лексикографии, возможно кратко отметить, что лексикография, а точнее, Канцелярия Словаря чешского языка стояла у истоков нынешнего Института чешского языка в Праге, который в 2011 году отметил 100 лет со дня основания. Институт языкознания им. Людовита Штура САН в Братиславе тоже будет праздновать круглую дату – 70 лет от основания тогда ещё Института языкознания Словацкой Академии наук и искусств, которое состоялось 1 апреля 1943 года. На заре его деятельности основной задачей была подготовка „Правил словацкой орфографии“ (изданы в 1953 году). Словацкая академическая лексикография начала развиваться в ИЯЛШ специально в рамках создания толкового словаря среднего типа – известного „Словаря словацкого языка“ в 6-ти томах под редакцией Ш. Пецьяра, который начали составлять и издавать позже (1959–1968). Лексикографические труды различного характера с тех пор стали и, пожалуй, остаются одной из самых важных сфер деятельности института на протяжении всех семи десятилетий его существования. Их составлением занимались сотрудники всех основных отделов института. Постепенно были опубликованы словари:

– **переводные**: „Чешско-словацкий словарь“ (Česko-slovenský slovník. Red. G. Horák. Bratislava, Veda 1979 (1-е изд.), 1981 (2-е изд.)), „Большой русско-словацкий словарь“ (Veľký rusko-slovenský slovník. Red. J. Horecký (1 и 2 том), D. Kollár (3–5 том). Bratislava, Vydavateľstvo SAV 1960–1970), „Большой словацко-русский словарь“ (Veľký slovensko-ruský slovník. 6 zv. Red. E. Sekaninová. Bratislava, Veda 1979–1995);

– **исторический**: „Исторический словарь словацкого языка“ (Historický slovník slovenského jazyka. 7 zv. Red. M. Majtán. Bratislava, Veda 1991–2008);

– **диалектный**: „Словарь словацких диалектов“ (Slovník slovenských nárečí. Bratislava, Veda 1994 (1 том), 2005 (2 том));

– **словарь синонимов**: „Словарь синонимов словацкого языка“ (Synonymický slovník slovenčiny. Red. M. Pisárčiková. Bratislava, Veda 1995 (1-е изд.), 2000 (2-е изд.), 2004 (3-е изд.));

– **иностранных слов**: „Словарь иностранных слов“ (Slovník cudzích slov (akademický). Bratislava, Slovenské pedagogické nakladateľstvo 1997 (1-е изд.), 2005 (2-е изд.)) и, конечно,

– **толковые**: „Краткий словарь словацкого языка“ (Krátky slovník slovenského jazyka. Red. J. Kačala – M. Pisárčiková (1-е – 3-е изд.), J. Kačala – M. Pisárčiková – M. Považaj (4-е изд.). Bratislava, Veda 1987 (1-е изд.), 1989 (2-е изд.), 1997 (3-е изд.), 2003 (4-е изд.)) и прежде всего новейший „Словарь современного словацкого языка“ (Slovník súčasného slovenského jazyka. Red. K. Buzássyová – A. Jarošová. Bratislava, Veda 2006 (1 том A – G), 2011 (2 том H – L)).

Братиславская академическая лексикография приобрела известность не только в славянской и не только в лингвистической среде. Она имеет свои традиции, теоретическую базу, концепции и преемственность, и, хотя словацкий язык относится к т. н. малым языкам с относительно небольшим числом носителей, он обработан практически во всех областях (кроме этимологии, но и здесь работа над словарём уже близится к завершению). Действительно широкому применению лексикографических трудов в исследовательской, научной и обиходной практике помогли компьютеризация указанных работ и доступ к некоторым из них в интернете http://slovniky.korpus.sk/), где мы регистрируем в среднем 40 000 запросов в день. Остальные словари также готовятся к размещению в интернете, пока они доступны только в локальной сети ИЯЛШ, где служат в качестве лексической базы данных как один из основных источников языкового анализа или как дополнительный источник при составлении следующих томов (словарь современного языка, диалектный словарь) и новых словарей. В 2011 году в базу данных в интернете были добавлены все 5 томов словаря А. Бернолака из 1825 г. (A. Bernolák: Slowár Slowenskí Češko-Laťinsko-Ňemecko-Uherskí, http://www.juls.savba.sk/ediela/bernolak/). Ценным дополнением к лексикографическому описанию современного словацкого языка являются специализированные словари чрезвычайно продуктивных авторских коллективов из Прешовского университета: словарь словосочетаний, морфемный словарь, словарь корневых морфем (подробнее в списке литературы, см. и Šimková 2008a); ещё несколько находятся в стадии разработки.

Важнейшим источником материала для лексикографических описаний современного словацкого языка уже с середины 1990-х годов стали тексты, обработанные в электронном виде в рамках баз данных корпуса. Сначала это был корпус текстов словацкого языка для внутреннего пользования ИЯЛШ, составленный в основном добровольцами и довольно несистематически в процессе выполнения других задач без соответствующего технического обес-

печения, поэтому его объём был невелик (за шесть лет – менее 30 миллионов текстовых слов), а тексты практически не снабжались лингвистическими аннотациями. В 2002 году возник самостоятельный отдел Словацкого национального корпуса ИЯЛШ САН (http://korpus.juls.savba.sk), главной задачей которого была и есть электронная обработка словарного запаса современных письменных словацких текстов, обиходного устного словацкого языка, а также всех лингвистических данных и обеспечение доступа к ним в интернете в научно-исследовательских и учебных целях.

В то время как „докорпусные“ лингвистические исследования проводились на материале, собираемом путём выборки в форме бумажной картотеки (напр., общая картотека ИЯЛШ САН насчитывает свыше 5 миллионов карточек) и на основании языкового сознания и интроспекции исследователей, современные корпусные базы данных достигают объёма в сотни миллионов и несколько миллиардов единиц. В Словацком национальном корпусе за десять лет существования было создано несколько текстовых баз данных, которые в настоящее время содержат около 2,5 миллиардов единиц-токенов (слов, знаков препинания, различных символов и находящихся в реальных текстах знаков). Тексты оцифровывали из классического печатного варианта, получали прямо в электронной форме из издательств или от авторов, а в случае устного корпуса переписывали с аудиозаписей. Основной корпус письменных текстов содержит в своей нынешней версии Prim-6.0 почти 1,2 миллиарда токенов (http://korpus.sk/stats.html). Из его общедоступной версии выделяются подкорпусы для целевого анализа и применения в специальных лексикографических целях: публицистический подкорпус, научный и научно-популярный подкорпус, подкорпус художественных текстов, подкорпус оригинальных произведений словацкой литературы, подкорпус текстов с 1955 по 1989 год. Базы данных письменных текстов включают специализированный корпус юридических текстов legal-1.0 (почти 150 миллионов токенов) и параллельные корпусы (http://korpus.sk/extra.html_): словацко-французский, словацко-русский, словацко-чешский, словацко-английский, словацко-латинский, готовятся словацко-болгарский, словацко-венгерский и словацко-немецкий параллельные корпусы. Для исследования стандартной формы словацкой разговорной речи с 2008 г. формируется Словацкий разговорный корпус (http://www.korpus.sk/shk.html), четвёртая версия которого в настоящее время насчитывает более 2,6 млн. токенов. Отдельным источником стал в последнее время и интернет, из которого словацкие тексты собираются в специальный веб-корпус (версия 2.0 содержит 1,1 мрд. токенов).

Каждая база данных доступна в интернете бесплатно для всех желающих, работать с ними можно через приложение Bonito (1-й и 2-й версиях), для внутренних нужд ИЯЛШ, в частности, для лексикографических целей, используется Sketch Engine (http://www.sketchengine.co.uk/). Весь материал структурирован, лингвистически аннотирован, поиск в нём можно производить, кроме слов, также на основании грамматических категорий слов и

словосочетаний (часть речи, число, падеж, время, вид и т. п.) и на основании внешних признаков текста (стиль, жанр, год издания, оригинальный словацкий или переводной текст и т. п.). В такой форме это уже не просто набор текстов, а комплекс языковых данных и лингвистических источников, составной частью которых является морфологическая база данных, содержащая почти 100 тыс. слов с более чем 4,5 млн. форм. Базу данных можно использовать, например, в обучении словаков и иностранцев словацкому языку, однако изначально она была создана для автоматической аннотации корпуса и компьютерной обработки словацкого языка. (Более подробно и о технических процедурах создания корпусов и их структуре см. Šimková, Garabík 2012). В настоящее время добавляются источники нелитературных форм национального языка — так, был добавлен исторический корпус словацкого языка (http://korpus.sk/extra.html), готовится корпус диалектный.

Создание Словацкого национального корпуса и лексикальных баз данных изначально предполагалось в первую очередь для составления словарей, поэтому методология его составления и сейчас отражает прежде всего лингвистические потребности (см. Šimková 2008b). Члены лексикографического коллектива регулярно обеспечивают обратную связь относительно его структуры, аннотирования и программного обеспечения и формулируют свои требования по дополнению и улучшению корпусных баз данных. Можно сказать, что в последние двадцать лет для словацкой лингвистики является характерным взаимный симбиоз, с одной стороны, классической и компьютерной лексикографии, а с другой, компьютерной обработки словацкого языка, применения компьютерно-лингвистических методов, создания корпусов и применения корпусно-лингвистических методов. Насколько нам известно, столь кооперативно и длительно эти процессы не протекают в других славянских языках. В эти процессы включаются также грамматические исследования, правда, пока лишь фрагментарно (несколько отдельных проектов и индивидуальных исследований в Братиславе и Прешове; наибольшее внимание грамматистов при работе с корпусами и языковыми источниками традиционно обращается на существительные и глаголы либо избранные категории и проблемы существительных и глаголов).

Какой опыт дал нам такой симбиоз создателей корпусов и создателей словарей и как он может развиваться дальше?

1. Ответ на вопрос, что же было раньше – язык, корпус или словарь, и что чем детерминировано, довольно прост: если нет языка и хотя бы основ его грамматического познания, не может быть словаря, однако без корпуса и без большой базы материала в любой форме (даже без компьютерной поддержки) словарь возникнуть может, и такие словари в прошлом нам известны. В Словакии до начала 1990-х гг. – до эпохи корпусной и компьютерной лингвистики и компьютерной лексикографии – было составлено много хороших словарей, которые служат обществу и лингвистике и по сей день. Правда, со второй половины прошлого века лексикографы уже имели в своем распоряжении систематически создаваемую общую картотеку и несколько

специализированных картотек, сознавая, что работа лексикографа не может обойтись без теоретической базы и материала. И совершенно справедливо требовали более современного материального, а с развитием компьютерной техники и компьютерного оснащения, поскольку не только профессиональные пользователи, но и любители ожидают от них высококачественных словарей с объективным и подробным описанием максимального объема языковых средств – описанием, опирающимся на стройную теорию и сделанным на основе аутентичных текстов различных стилей и жанров в рамках всего языкового сообщества.

2. Зачем же нужны корпусы?

Этот вопрос постоянно задавался на начальном этапе создания электронных баз данных, возникавших, во-первых, для нужд (прежде всего английской и американской) лексикографии. Особенно бурные дискуссии приводили к обогащению текстов и текстовых слов лингвистической информацией (морфологическая, синтаксическая, семантическая и др. аннотации). Между тем развитие компьютерной техники и корпусных инструментов, а вместе с ним и (корпусно-)лингвистического мышления постепенно давало возможность приемлемого для всех аннотирования: текст/текстовые слова можно использовать/анализировать как чистый материал без внешней или внутренней языковой аннотации; аннотацию можно корректировать или дополнять при одновременном сохранении или удалении первичной и т. п. Корпусы с целым комплексом данных и инструментов развились в блоки языковых источников, и специалисты перешли от вопроса „Для чего нужен корпус?“ ко все более глубокому пониманию того, что это источник действительно объективного познания языка и реального функционирования языковых средств. И возникали вопросы:

→ „Как построить корпус?“ (в его составе концепция репрезентативности/пропорциональности национального корпуса, особенности создания корпусов письменных/устных текстов, одноязычных/параллельных корпусов и др.);

→ „Как построить хороший корпус?“ (касается объема и качества текстов и аннотаций, учета потребностей лингвистов, специалистов по компьютерной обработке естественного языка и др.);

→ „Какие корпусы необходимо/полезно еще составить?“ (синхронные/диахронные, литературные/нелитературные, классические/интернет-корпусы...);

→ „Какие инструменты создать для более эффективной работы с корпусами?“ (исключение избыточного материала, случайный выбор, отбор и сортировка материалов, статистические инструменты);

→ „Как выбрать подходящий корпус для конкретного исследования?“ (особенно важный вопрос при сравнительных исследованиях различных языков, корпусы которых отличаются по объему и созданы различными методами, что оказывает влияние на итоговые статистические данные).

3. Для чего нужна компьютерная лексикография?

Со времени изобретения книгопечатания итоговая обработка текстов постоянно совершенствовалась: от ручной к механической, машинной, полуавтоматической и вплоть до компьютерной типографии, т.е. окончательной технической и графической подготовки текста к печати. Составление словарей не могло остаться в изоляции от этого развития, и от рукописного составления лексикографы перешли к машинописному, полукомпьютерному (перенос готовых словарных статей в компьютер с последующей компьютерной обработкой – сортировка, типография) и, наконец, к полностью компьютерному составлению, редактированию и корректурной правке. Лексикографические коллективы в Словакии (в Братиславе и Прешове) имеют собственные лексические базы данных, в которых можно искать и сравнивать словарные статьи опубликованных или разработанных словарей по разным критериям (количество значений, квалификаторы, языковое происхождение, диалектная зона и т. п.), новые статьи вносятся непосредственно в компьютер в специальном режиме редактирования, регулярно проверяются, технически корректируются, унифицируются, оценивается единство обработки лексико-семантических групп, единство оформления ссылок и т. д. Компьютерная лексикография, или как минимум компьютерная поддержка лексикографии, развивалась благодаря все большей интернационализации лексикографической работы (возникновению и деятельности различных лексикографических ассоциаций, конгрессов, изданий – см. и Jarošová 1997), которая продолжает интернационализироваться, а ее качество повышается благодаря международным проектам, направленным на построение лексикографической инфраструктуры, различных транснациональных лексикографических сетей, что, в свою очередь, обогащает также национальные лексикографические традиции.

4. Как к этому относятся опытные/традиционные лексикографы (и грамматисты)?

По результатам некоторых анализов и по своему опыту мы знаем возможности и ограничения корпусов, корпусной лингвистики и компьютерной лексикографии (ср. напр. Štícha 2006; Sokolová, Šimková, Ivanová 2006; Cvrček, Kováříková 2011). Ограничения, к сожалению, начинаются уже во внеязыковой сфере при обычном отключении электричества или одного из серверов, при плохо продуманной системе сохранения резервных копий и/или безопасности компьютерных баз, а продолжаются при абсолютизировании технического подхода, небрежной работе с частотными характеристиками, при ожидании того, что из корпусов словарь с помощью компьютерной лексикографии составится сам... Очевидно, что и при достаточном количестве, а иногда и избытке языкового материала и компьютерных инструментов первое и последнее слово остается за человеком, который должен высидеть и выработать все это. И хотя материал и инструменты открывают перед ним ранее немыслимые возможности, дают новый взгляд на язык и его функционирование, они в то же время заставляют его переоценить ранее проверенные

методы. Как указывает А. Ярошова, одна из ведущих современных словацких лексикографов, с опорой на Дж. Синклера и других специалистов в данной области, „данные о функционировании слова в конкретных текстах, которые корпус предоставляет в доселе невиданном количестве, изменяют и наш взгляд на существо значения слова и требуют переоценки существующей модели значения“ (Jarošová 2003: 60), переоценки типов словосочетаний, их принадлежности к языку (langue) или речи (parole), удачности эквивалентов, обоснованности словоцентрической позиции и полезности текстоцентрической (коммуникативной) позиции при оценке эквивалента в двуязычной лексикографии и т. п. (там же, с. 61 и сл.).

Развитие корпусной лингвистики и компьютерной лексикографии является необратимым, а технические возможности лексикографов и объем материала с каждым днем расширяются. Довольно высокая компьютерная грамотность словацких лексикографов и хорошая материально-техническая база института (лексические базы данных, Словацкий национальный корпус, Bonito 1, Bonito 2, Sketch Engine), несомненно, помогли повышению качества и эффективности лексикографического труда. Однако ожидаемого ускорения лексикографического труда (в словацкой среде – ожидаемое издание одного тома большого толкового словаря каждый год), вопреки всем позитивным моментам и хорошим предпосылкам, пока не произошло: материала часто оказывается слишком много, хотя его бывает недостаточно по некоторым жанрам и терминологическим областям или для документирования окказиональных слов, словоформ и словосочетаний; язык и его познание постоянно развиваются, поэтому и материал динамически пополняется, улучшаются его качество, аннотирование и компьютерные инструменты его обработки, поэтому лексикограф должен быть не менее динамичным. В словацких условиях лексикографам не хватает актуальных грамматических и орфоэпических источников, последнее кодифицированное издание орфографических норм вышло 15 лет назад (в данный момент последние изменения в правилах словацкого правописания содержит издание Pravidlá slovenského pravopisu 1998), поэтому очень много времени затрачивается при составлении соответствующих частей словарных статей и ведется много дискуссий.

5. Каковы перспективы дальнейшего развития корпусов, корпусной и компьютерной лингвистики и компьютерной лексикографии?

Пока жив язык и его носители/пользователи, словари будут необходимы. Для их качественного составления непременно понадобится ноу-хау целой области компьютерной обработки естественного языка, включая компьютерную лексикографию и материал, который в достаточном объеме и на адекватном развитию человечества и его знаний уровне смогут обеспечить именно корпусы. Составление текстовых корпусов с самого начала руководствовалось принципом „чем больше (материала), тем лучше“, и этот принцип в соответствии с развитием компьютерных технологий применяется по сей день. Кроме интересов лингвистики и компьютерной обработки естествен-

ного языка, он поддерживается и вопросами простых носителей языка, напр.: „Сколько слов в словацком языке?“, „Почему в корпусе есть не все тексты?“. Иногда встречаются и представления вроде „Чего нет в корпусе, того нет и в языке“ или „Когда корпус будет вмещать весь язык?“. Вопреки значительному прогрессу и постоянному развитию в этих областях, скорее всего, никогда не удастся зафиксировать весь спектр функционирования языка в электронных базах данных, хотя и очевидно, что корпусы приближаются к фиксации языка как целого вместе с его актуальной динамикой более всего, что было раньше. Если словарь всегда представляет лишь малый фрагмент языка и знаний о нем в соответствии со своим типом, причем фрагмент, как правило, на несколько лет отстающий от актуального состояния языковой действительности (словари обыкновенно составляются на протяжении довольно значительного времени), корпус представляет несравнимо больший фрагмент, постоянно приближающийся к целому языку, а отставание от актуального состояния языка сокращается до нескольких месяцев (напр. в Словацком национальном корпусе основной корпус письменных текстов обновляется каждые два года, а при выходе в свет новой версии корпус уже содержит материал, опубликованный всего за несколько недель до появления нового корпуса в интернете).

Заключение. Современная лингвистика смещается по направлению от теоретической схемы языка, создаваемой на основании нескольких избранных явлений, в область эмпирических наук, чему не в последнюю очередь способствуют корпусная и компьютерная лингвистика со своей стратегией „как можно больше как можно лучших данных“. Корпусные и компьютерные отделы обычно находятся в компьютерных и математических институтах, но в словацких условиях подтвердилась необходимость создания самостоятельного отдела Словацкого национального корпуса в Институте языкознания им. Людовита Штура Словацкой Академии наук, где составляется новый толковый словарь – „Словарь современного словацкого языка“ – и готовятся другие лексикографические и грамматические описания словацкого языка. Сотрудничество и взаимодействие создателей и пользователей компьютерных баз данных и инструментов, как показывает наш опыт, оказалось исключительно полезным для обеих сторон.

ЛИТЕРАТУРА

Cvrček, Kováříková 2011: Cvrček, V., D. Kováříková. Možnosti a meze korpusové lingvistiky. // *Naše řeč*, č. 3, s. 113–133.

Jarošová 1997: Jarošová, A. Lexikografia a počítače – slovenský variant. // *Sociolinguistica Slovaca*, 3, s. 304–311.

Jarošová 2003: Jarošová, A. Impulzy korpusovej lingvistiky pre dvojjazyčnú lexikografiu: medzi textovou a slovníkovou ekvivalenciou. // *Jazykovedný časopis*, č. 1–2, s. 59–68.

Nižníková, Sokolová a kol. 1998: Nižníková, J., M. Sokolová a kol. *Valenčný slovník slovenských slovies*. 1. zv. Prešov: Filozofická fakulta Prešovskej univerzity. 270 s.

Nižníková, Sokolová a kol. 2006: Nižníková, J., M. Sokolová a kol. *Valenčný slovník slovenských slovies*. 2. zv. Prešov: Filozofická fakulta Prešovskej univerzity. 166 s.
Sokolová, Moško, Šimon, Benko 1999: Sokolová, M., G. Moško, F. Šimon, V. Benko. *Morfematický slovník slovenčiny*. 1. vyd. Prešov: Náuka. 530 s.
Sokolová, Ološtiak, Ivanová, Šimon, Czéreová, Vužňáková, Benko, Moško 2005: Sokolová, M., M. Ološtiak, M. Ivanová, F. Šimon, B. Czéreová, K. Vužňáková, V. Benko, G. Moško. *Slovník koreňových morfém slovenčiny.* 1. vyd. Prešov: Filozofická fakulta Prešovskej univerzity. 584 s.
Sokolová, Ološtiak, Ivanová a kol. 2007: Sokolová, M., M. Ološtiak, M. Ivanová a kol. *Slovník koreňových morfém slovenčiny*. 2., upravené vydanie. Prešov: Filozofická fakulta Prešovskej univerzity. 586 s.
Sokolová, Ološtiak, Ivanová a kol. 2012: Sokolová, M., M. Ološtiak, M. Ivanová a kol. *Slovník koreňových morfém slovenčiny*. 3., upravené a doplnené vydanie. Prešov: Filozofická fakulta Prešovskej univerzity. 690 s.
Sokolová, Šimková, Ivanová 2006: Sokolová, M., M. Šimková, M. Ivanová. Možnosti a medze lingvistického výskumu v Slovenskom národnom korpuse. // *Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli*. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove, s. 7–14.
Šimková 2008a: Šimková, M. Korpusová lingvistika na Slovensku. // *Jazykovedný časopis*, č. 1–2, s. 11–24.
Šimková 2008b: Šimková, M. Морфологическая разметка частей речи в Словащком национальном корпусе и возможности её использования в процессе создания танкового словаря. // *Труды международной конференции „Корпусная лингвистика“*. Санкт-Петербург: Изд. Санкт-Петербургского университета, с. 387–395.
Šimková, Garabík 2012: Šimková, M., R. Garabík. The Slovak National Corpus and its Corpus Linguistic Resources. // *Prace filologiczne*. Tom LXIII. Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego, s. 109–119.
Štícha 2006: Štícha, F. (ed.) *Možnosti a meze české gramatiky*. Praha: Academia. 304 s.