Репрезентативность корпуса как лингвистическая проблема

Мария Шимкова

В лингвистических исследованиях в разных фазах развития неизбежно решался вопрос репрезентативности материала, на базе которого создавались лексикографические труды и грамматические справочники. Например, уже академик А. А. Шахматов особо подчеркивал, что основной задачей описания современного русского языка является собирание и систематизация «надежного материала, из которого было бы видно, как говорит народ в различных областях России, как выражаются современные писатели, в каком значении употреблялись те или другие слова писателями прежнего времени и т. д.» [1, с. 31]. Очевидно, что эту задачу сегодня успешно выполняют корпуса, которые заменяют предшествующие картотеки. Обе формы накопления и систематизации свидетельств о функционировании лингвистических единиц в контекстах должны иметь какие-либо определенные критерии построения. В предшествующих периодах исследования, в т. ч. и картотеки, ориентировались на т. н. репрезентативный язык, т. е. язык «высокого» стиля выдающихся авторов оригинальной литературы. Сегодня понятие репрезентативность текстов имеет в корпусе другое содержание и решается циклически от начала создания и использования корпусов в отдельных языках. Об этом свидетельствуют и некоторые статьи в сборнике с конференции «Корпусная лингвистика – 2004» [2, 3].

Полагаем, что вопрос репрезентативности корпуса – прежде всего лингвистическая проблема, которая решается на различных, в т. ч. и экстралингвистических, уровнях: объем, стили и жанры, временна́я перспектива, географическая перспектива (регионы, издательства), лингвистические единицы. В статье предлагаем опыт, приобретенный при создании и использовании Словацкого национального корпуса (СНК).

СНК основывается на традиции предшествующего корпуса текстов

словацкого языка, который создавался в Институте языкознания им. Л. Штура Словацкой академии наук в Братиславе практически с 1993 по 2002 гг. как электронная база данных лингвистически необработанного текстового материала, прежде всего для нужд лексикографического коллектива, составляющего новый толковый словарь современного словацкого языка. СНК, в определенной степени, продолжает ориентироваться на пользователя-лексикографа, но распространил свои возможности и на обычного пользователя (любители, интересующиеся языком, студенты, учителя, редакторы и другие, работающие со словом и текстом), а также на специалистов в области грамматических исследований и компьютерной обработки естественного языка.

Каждая категория пользователей имеет собственные потребности. Обычные пользователи чаще всего ищут в корпусе слово или словосочетание, когда сражаются с какой-либо орфографической или стилистической проблемой (написание заимствованных слов, имен собственных, поиск синонимов, значений новых слов и т. п.). Для этих нужд подходит корпус текстов в лингвистически необработанной форме и такого размера, чтобы покрыл общеупотребительный словарный запас, включая новые слова, и общеупотребительные языковые средства.

Более высокие требования к корпусу имеет пользователь-лексикограф, который приветствует как можно более расширенную базу данных корпуса (more data is better data), покрывающую не только обычный словарный запас, но содержащую редкие слова и языковые средства. Чем больше разрастается корпус, тем очевиднее становится необходимость автоматизации некоторых его процедур. Большой объем текстов в корпусе одновременно приносит не очень желаемый результат в форме многотысячных (иногда и миллионных) списков обычных слов или языковых средств. Для работы, и в этом случае, может быть достаточно корпуса текстов без дополнительной лингвистической информации, но все-таки необходимыми являются данные библиографического характера и,

иногда, информация о стиле и жанре.

Существенно более требовательной группой являются потенциальные пользователи из среды лингвистов-теоретиков и специалистов в области компьютерной обработки естественного языка, для которых необходим лингвистически размеченный корпус с соответствующими данными лексикально-семантического, морфологического, синтаксического характера и т. п. Создатель корпуса должен принимать во внимание все эти требования и особое внимание уделять репрезентативности корпуса. На этот вопрос, как уже упоминалось в начале, неоднократно обращали внимание [4, 5].

Корпусовые статистики сообщают, что в корпусе, который содержит 100 миллионов текстовых единиц, что в настоящее время является минимальным размером обычного корпуса, 8 тыс. единиц находятся в 95% текста, а остальные 5% представляют 500 тыс. единиц. Это можно решить все большим корпусом или, кажется, созданием репрезентативного корпуса, который бы включал в себя тексты, как можно шире охватывающие стили и жанры, поколения и группы авторов, издательские практики и т. п., а также содержащие целую шкалу языковых средств.

При рассмотрении репрезентативности необходимо определить целостность, по отношению к которой корпус должен быть репрезентативен. Если говорить о всеобщем национальном корпусе, то было бы правильно предположить, что он должен был бы включать все, что когда-либо в данном национальном языке встречалось. Немало современных корпусов этого типа (включая СНК) находятся на уровне письменных текстов, преимущественно синхронных. Компьютерная обработка спонтанных устных проявлений, а также диахронных текстов все еще является экономически трудновыполнимой. Всеобщий корпус, нацеленный на пропорциональный охват языковой практики, по анализам Д. Байбера [4, с. 116], должен был бы содержать приблизительно 90% разговоров (обычной разговорной речи), 3% писем и замечаний и 7%

опубликованных и неопубликованных текстов классических стилей и жанров. Это практически наоборот, как мы сейчас делаем.

И если всеобщий национальный корпус составляется из письменных текстов современного языка, при создании проекта все же необходимо с хронологической точки зрения ограничить современный язык, и с точки зрения репрезентативности определить, будут ли тексты для корпуса подбираться на основе принципа адекватного представления всех стилей и жанров (типов текстов) или на основе адекватного размещения языковых явлений в соответствующих текстах / целом корпусе (например, в публицистических текстах можно ожидать, что я-формы будут встречаться довольно редко, а в специальных текстах технической направленности – бедность синтаксических структур и т. п.). Главным образом второй принцип предполагает проведение эмпирического исследования на обширном, лингвистически обработанном материале, т. е. при помощи лингвистически размеченного корпуса.

В случае принятия первого принципа решаются вопросы критериев для деления текстов. Д. Байбер [4] предлагает использовать для иерархизации уровней подбора и для распределения текстов по «регистрам» ситуационные параметры, причем регистр понимает как понятие скорее непрерывное, а не дискретное:

1. первичный канал – письменная речь / устная речь / записанная устная речь;
2. формат – опубликованный / неопубликованный;
3. сцена – в рамках учреждения / другая общественная / частная, личная;
4. адресат – а) множественность: (не)названный / коллективный / индивидуальный / автор как адресат; б) присутствие (место и время): присутствующий / отсутствующий; в) интерактивность: отсутствует / незначительная / значительная; г) знание: общее / специализированное / личное;
5. адресор – а) демографическая вариация: пол, возраст, род занятий и др.; б) признание благодарности: указанная персона / учреждение;
6. фактуальность – фактуально-информационная / средняя или неопределенная / воображаемая;
7. назначение – убедить, развлечь, возвысить, информировать, научить, объяснить, говорить, описать, записать, самовыразиться, выразить отношение, мнение или эмоции, укрепить межчеловеческие отношения...
8. темы...

Деление текстов на основе приведенных характеристик значительно

отличается от традиционной словацкой стилево-жанровой классификации, особенно два последних параметра требуют дальнейшего теоретического и эмпирического изучения. Несмотря на это, возможность стратификации текстов по регистрам, используя ситуационные параметры, представляет, с определенной точки зрения, более простой способ при установлении репрезентативности корпуса. Однако идеальный корпус должен представлять не только шкалу регистров, но и шкалу разнообразности, распределения языковых средств в отдельных типах текстов. По мнению Д. Байбера [4, с. 118 – 127]:

а) повседневные линейные языковые явления распределяются в текстах сравнительно стабильно, и их можно достоверно установить в относительно коротких текстовых сегментах (уже в объеме 1000 слов);

б) редкие языковые явления обнаруживают значительное разнообразие распределения и требуют более объемные образцы текстов;

в) явления с распределением вероятности по кривой, т. е. различные типы явлений (например, накопляемость частей речи) – относительно стабильны во всех следующих один за другим текстовых сегментах, но число появлений новых типов в тексте постепенно понижается, причем частота новых типов является во всех текстовых сегментах систематически более высокой, чем в образце одного текста. Иными словами, больше разнообразия в типах текстов, включенных в корпус, спроецируется на более широкую репрезентативность типов языковых явлений, причем образцы текстов должны быть достаточно объемные, чтобы могли достоверно представлять распределение языковых явлений.

Различные проекты корпусов подходят к вопросу подбора текстов и репрезентативности корпуса по-разному: от абстрагирования от этого свойства корпуса (Bank of English, Mannheimer Corpora) до детальной проработки процентного представления стилей и жанров на основе обширных и повторяемых социолингвистических исследований (Чешский национальный корпус). Многие корпусовые проекты постепенно

вырабатывали собственные критерии построения корпуса [6, с. 21 – 22], которые в настоящее время могут объективироваться на основе стандартизированных рекомендаций TEI (Text Encoding Initiative) и EAGLES (Expert Advisory Group on Language Engineering Standards). Общей методикой, исходя из предыдущего опыта, является, прежде всего, цикличность составления репрезентативного корпуса: по теоретическому определению ситуационных параметров, детерминирующих выбор текстов в данной языковой общности, и по определению объема важных языковых явлений, которые будут в корпусе анализироваться, создается пилотный корпус с достаточно широкой шкалой разнообразности, и с глубиной текстов и регистров. Эти тексты будут грамматически размечены, и на пилотном корпусе будет осуществлено эмпирическое исследование, результаты которого подтвердят или изменят примененные параметры. Отдельные фазы этого цикла должны протекать безостановочно и динамично модифицировать общий характер корпуса.

Такой прием, например, используется в Чешском национальном корпусе, где созданию первого репрезентативного корпуса чешского языка SYN2000 предшествовало предварительное исследование, на основе которого было определено представление главных стилей в таком объеме: художественная литература и литература факта – 15%, газеты и журналы – 60%, специальная литература – 25%. Большая доля газет и журналов исходит из первоначального исследования Opinion Window Prague 1996. Но обнаружилось, что с тех пор, во-первых, изменились условия издания и, следовательно, читаемость периодик (уменьшилось количество журналов и ежедневных газет, их тиражи, и выросли цены), во-вторых, в 2001 г. проводились два новых исследования на основе другой методики и с помощью по-разному поставленных вопросов. Результатом является предложение новой структуры Чешского национального корпуса в форме: 40% – художественная литература и литература факта, 33% – газеты, и журналы, 27% – специальная литература. Административные тексты,

которые были при первоначальном распределении включены в специальные тексты, по новой методике вообще не попадают в корпус из-за их информационной ценности, которая, по существу, такая же, как у специальных текстов; пользователи работают с ними так же, как с инструкциями, рекомендациями, правилами, и не читают их систематически, как другие виды литературы [7].

При подготовке проекта СНК мы исходили из опыта существующих корпусовых проектов, главным образом чешских, из потребностей потенциальных пользователей электронной базой данных словацких текстов и из реальных возможностей небольшого коллектива, строящего корпус небольшого языка. В рамках подготовки концепции на 2003 – 2006 годы [8, 9] мы установили в качестве приоритета создание всеобщего одноязычного корпуса письменных текстов современного словацкого языка (1955 – 2005). При сборе данных мы сначала решали вопрос представляет ли идеально репрезентативный корпус действительно реальное функционирование языковых единиц, или их какую-то репрезентативную картину, и потом руководствовались принципом «как можно больше и как можно более разнообразных текстов». О репрезентативном образце письменных текстов современного словацкого языка мы размышляли в общих чертах: 1/3 публицистических текстов, 1/3 художественных текстов и 1/3 специальных и научно-популярных текстов. (Бесспорным подтверждением правильности этого первичного основного распределения являются и уже упомянутые чешские социолингвистические исследования, и предложение новой структуры Чешского национального корпуса). В последних двух группах мы делали упор на переводы, которые имеют в меньших национальных и языковых общностях, какой является и словацкая общность, особое положение, но при подготовке предшествующих лексикографических справочников словацкого языка никак не были представлены. Для СНК мы предлагали приблизительно третью часть переведенных художественных и специальных (или научно-

популярных) текстов. Среди публицистических текстов тоже встречаются переводы, но их почти невозможно определить (переводы новостей, предоставленных агентствами, без ссылки на то, что это перевод, и эту информацию невозможно автоматизировано уловить).

СНК в настоящее время доступный в Интернете (http://korpus.juls.savba.sk) с сегментациейи стилево-жанровой аннотацией в версии prim1 (примарный корпус), которая содержит почти 200 млн. единиц. Структура этих данных представляет почти 95% публицистических текстов, 3,5% художественных текстов и 1,5% специальных и научно-популярных текстов, изкоторых большая часть (98%) с 1994 по 2004 гг. Заметна диспропорция в пользу публицистических текстов, поэтому приступили к созданию тестирующей версии сбалансированного корпуса. Принимая во внимание малое количество специальной литературы, было возможно создать такой корпус только в объеме 12 млн. единиц в структуре 60% публицистики, 20% художественной и 20% специальной и научно-популярной литературы. На таком корпусе можно проводить, и уже проводятся, эмпирические исследования, ведущие к созданию репрезентативного корпуса письменных текстов современного словацкого языка. Установленная частота текстовых единиц (лемм) настоящей структуры текстов в СНК уже сейчас обнаруживает стандартное размещение наиболее часто встречающихся предлогов, союзов, местоимений и частиц, даже и существительных как известно из предшествующих исследований [10]. Практически те же самые часто встречающиеся языковые единици показываются в репрезентативном Чешском национальном корпусе. При отборе текстов для морфологической и синтаксической разметки вручную мы также принимали во внимание необходимость разнообразия. В первой фазе была сделана аннотация к одному художественному переводу, и за отобранными текстами из интернетжурнала InZine следовали тексты из специальной и научно-популярной литературы.

СНК предоставляет в сегодняшнем состоянии основной исследовательский материал для всех категорий пользователей. Но он не заменяет орфографические и грамматические справочники; представляет лишь исходную точку для их создания, но исходную точку с хорошей доступностью в рамках работы с интернетресурсами и с большими возможностями в рамках автоматизированной обработки большого количества реальных текстов. После окончания работы над указанными задачами СНК будет постепенно выполнять критерии, возлагаемые на сбалансированный корпус национального языка общего типа. Как каждый такой корпус, он будет всегда отражать только реальные проявления языковых средств, находящихся в текстах, обработанных в корпусе, но не предоставит информацию о несуществующих языковых средствах, так как корпус может содержать большое количество материала, но вопреки любой удачной структуре не способен охватить всю языковую систему.

Литература

[1] Шахматов, А. А.: Несколько слов по поводу записки И. Х. Пахмана. Сборник ОРЯС, 1899, т. XVII, № 1, с. 31.

[2] Девель, Л. А.: Репрезентативность корпусов английского языка (данные учебных одноязычных словарей). Труды международной конференции «Корпусная лингвистика – 2004». Издательство С.-Петербургского университета 2004, с. 131 – 137.

[3] Шаров, С. А. - Савчук, С. О.: Типология текстов для представительного корпуса. Труды международной конференции «Корпусная лингвистика – 2004». Издательство С.-Петербургского университета 2004, с. 352 – 362.

[4] Biber, D.: Reprezentativnost v projektu korpusu. Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, c. 107 – 136.

[5] http://www.ilc.cnr.it/EAGLES96/texttyp.html

[6] Čermák, F.: Jazykový korpus: prostředek a zdroj poznání. Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, c. 15 – 37.

[7] Králík, J.: Aktualizace rozvržení zdrojů Českého národního korpusu s ohledem na revizi vyváženosti jeho struktury. Slovo a slovesnost, 65, 2004, č. 2, s. 133 – 141.

[8] Šimková, M.: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. Počítačová podpora prekladu. Zborník prednášok (Budmerice 22. - 23. máj 2003). Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2003, s. 15 – 19.

[9] Šimková, M.: Slovenský národný korpus – východiská a plány. Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 150 – 158.

[10] Mistrík, J.: Frekvencia slov v slovenčine. Bratislava: VydavateĽstvo Slovenskej

akadémie vied 1969. 726 s.

Summary

In a broad context, this paper describes a preparation, build-up and perspectives of the Slovak National Corpus. Especially the 200-million contemporary corpus of written Slovak and criteria for its building including question of the representation are describes.

national corpus, representation of corpora, annotation