ТОМСКИЙ ДИАЛЕКТНЫЙ КОРПУС: СБАЛАНСИРОВАННОСТЬ И РЕПРЕЗЕНТАТИВНОСТЬ

Земичева Светлана Сергеевна (optysmith@gmail.ru)

Национальный исследовательский Томский государственный университет, г. Томск, Россия,

The project of a new electronic resource – the search system created on the basis of the available dialect archive – is presented. The material for the corpus is records of dialect speech, which has been collected during 70 years in Tomsk and Kemerovo regions and translated into text format. The author describes types of the mark-up and search capabilities of the corpus. Characteristic of its representativeness and balance is produced on the background of other dialect corpora.

Keywords: corpus linguistics, Tomsk dialect corpus, Russian dialects of Siberia

Томский диалектный корпус, созданный в 2017 г., представляет собой новый источник исследования диалектной речи Сибири и традиционной культуры сибиряков [1].

Цель корпуса – облегчить трудоёмкий процесс сбора материала для диалектологов и лексикографов. Корпус находится в ограниченном доступе, что связано с необходимостью защиты персональных данных и неприкосновенности частной жизни информантов. Демонстрационная версия повторяет архитектуру основной части, но значительно меньше по объёму. Как и в Саратовском диалектном корпусе [2], принципиальной является ориентация на изучение не только собственно диалектной речи, но и традиционной народно-речевой культуры, репрезентированной в диалектном тексте.

Единицей выдачи в корпусе является текст, понимаемый как коммуникативный эпизод общения диалектоносителя с собирателями. Для каждого текста указываются место и время записи, архивный номер тетради (при наличии), сведения о говорящем – год рождения, уровень образования, информация о родителях и предках, о местах длительного проживания. Тексты подаются в упрощённой орфографической записи с сохранением отдельных фонетических особенностей.

На данный момент центральной является тематическая разметка. Её особенность в том, что тема маркируется не для текста в целом, а для каждого его фрагмента. Всего выделено 73 темы, представляющих зоны актуального внимания сельского жителя. Тематическая разметка носит уровневый характер, например, в составе макротемы «Природа» выделены темы «Местность», «Животные», «Растения», «Погода и атмосферные явления», «Стихийные бедствия», «Экология». В свою очередь, тема «Животные» членится на

подтемы «Дикие животные», «Домашние животные» и «Вредители». Такая классификация отражает наивное членение мира и отличается от строго научной таксономии.

Разметка по типам текста нацелена на разграничение текстов, близких к естественной коммуникации (диалог, полилог, ситуативное вкрапление) и «спровоцированных» (лингвистический опрос). Маркируется также фольклор как особый тип текста.

Предусмотрены следующие типы поиска: поиск слова, темы, типа текста; поиск по месту и году записи. Разные виды поиска могут комбинироваться между собой.

Информативные возможности корпуса расширяются благодаря включению ссылок на сканированные рукописей тетрадей (для старых записей) и мультимедиа данных: аудиофайлов, фотографий (для новых записей). Таким образом, корпус относится к типу мультимедийных.

Репрезентативность диалектного корпуса, если рассматривать её только как количественный показатель, определяется следующими параметрами: количеством обследованных регионов и населённых пунктов; количеством опрошенных информантов; количеством часов аудиозаписи; количеством зафиксированных словоупотреблений; временем наблюдения.

По этим параметрам Томский диалектный корпус характеризуется как достаточно представительный: на момент написания статьи в него входят материалы из 60 населённых пунктов Томской и Кемеровской областей, фиксирующие речь 800 информантов; представлены расшифровки и аудиофайлы в количестве 100 ч. записи; внесены материалы, собранные с 1947 по 2016 гг.

Общий объём корпуса на сегодня – 1 500 000 словоупотреблений[1].

Параметры репрезентативности диалектного корпуса, в частности, его объёма нуждаются в дополнительном теоретическом осмыслении. Если для национального корпуса считается достаточным объём в 100 млн. словоупотреблений, объём диалектного корпуса теоретически не определён.

Анализ существующих диалектных корпусов показывает, что обычно их объём составляет около 1 млн. словоупотреблений. Так, Хельсинкский корпус британских диалектов насчитывает 846 149 словоупотреблений [3], устная часть корпуса SCOTS

[1] На момент написания статьи для ввода в корпус подготовлено 20% рукописного архива и 70% расшифровок экспедиционных записей, сделанных с 2000-х гг. Общий объём архива составляет около 5 000 000 словоупотреблений, материалы ежегодно пополняются в ходе экспедиций.

включает 1,049,794 словоупотреблений [4], в эстонский диалектный корпус входит 1 284000 слов [5].

Наиболее репрезентативным из существующих русских диалектных корпусов является Устьянский корпус, его объём на март 2018 – 833 826 словоупотреблений [6]. Для сравнения: по состоянию на март 2018 г. в диалектный подкорпус НКРЯ входит 285 281 слово [7], в Ангарский корпус – 78 789 слов [8]. Небольшой объём корпуса в некоторых случаях компенсируется глубиной проработки текста, наличием сложной и многоуровневой разметки. Таков, в частности, диалектный подкорпус в составе НКРЯ [9].

Кроме того, если оценивать репрезентативность корпуса как модели реальной коммуникации по приближенности к последней, необходимо отметить, что записанные тексты являются «полуаутентичными» (включают как элементы спонтанной речи, так и материалы опросов, в том числе лингвистических интервью). Впрочем, аналогичная ситуация характерна для большинства диалектных корпусов. Для разграничения спровоцированных и спонтанных текстов в Томском диалектном корпусе проводится разметка по типу текста.

По параметру **сбалансированности** можно отметить следующее. Корпус является сбалансированным по территориальному признаку: разные группы говоров, выделенные в среднеобском диалекте (приобские, прикетские, притомские, нарымские) представлены достаточно равномерно, лишь группа причулымских говоров отражена несколько слабее остальных, т.к. включает некоторые нетипичные особенности.

В то же время корпус слабо сбалансирован по социолингвистическим параметрам: информанты относятся, в основном, к одной социальной группе – жителей села пенсионного возраста. Выбор информантов определялся научными интересами исследователей, поэтому в корпусе преобладают записи речи людей старшего поколения, хотя эпизодически отражена речь представителей других возрастных групп. Не сбалансирован корпус и в отношении половой принадлежности говорящих: 70% информантов – женщины, 30% – мужчины. Основная часть материала – речь сибирских старожилов, носителей среднеобских говоров, в меньшей степени отражается речь переселенцев, новосёлов. Достаточно сильно варьируется лишь уровень образования информантов: от полностью неграмотных (записи 60-х гг) до людей с высшим образованием (новейшие записи). Однако на данный момент сведения об образовании вносятся лишь в текстовое поле, поиск по ним невозможен. Представляется, что решение этой задачи в перспективе повысит информативные возможности ресурса.

Таким образом, созданный ресурс в целом является репрезентативным, а материалы сбалансированы по основному признаку – территориальному. Это позволяет считать его

достаточно надёжным источником для проведения диалектологических исследований и реализации лексикографических проектов.

Перспективы развития корпуса связаны с добавлением новых видов разметки и совершенствованием поисковых возможностей.

Библиографический список

1. Корпус: Демо-версия [Электронный ресурс] URL: http://losl.tsu.ru/?q=corpus_demo (дата обращения: 12.03.2018).
2. Крючкова О.Ю., Гольдин В.Е. Корпус русской диалектной речи: концепция и параметры оценки // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). М., 2011. С. 359–367.
3. *Helsinki Corpus of British English Dialects* [Электронный ресурс] URL: http://www.helsinki.fi/varieng/CoRD/corpora/Dialects/basic.html (дата обращения: 25.03.2018).
4. Scottish Corpus Of Texts & Speech [Электронный ресурс] URL: https://www.scottishcorpus.ac.uk/search/?search=Search&word=&author=&gender=-+All&region=-+All&spoken=y&title=&yearfrom=&yearto=&search=Search (дата обращения: 25.03.2018).
5. Estonian Dialect Corpus [Электронный ресурс] URL: http://www.murre.ut.ee/estonian-dialect-corpus/ (дата обращения: 25.03.2018).
6. Ustja River Basin Corpus [Электронный ресурс] URL: http://parasolcorpus.org/Pushkino/index.php (дата обращения: 25.03.2018).
7. *Диалектный* корпус национального корпуса русского языка [Электронный ресурс] URL: http://www.ruscorpora.ru/search-dialect.html (дата обращения: 12.03.2018).
8. *Электронный* текстовый корпус лингвокультуры северного Приангарья. Диалектный подкорпус. [Электронный ресурс] URL: http://angara.sfu-kras.ru/?page=dialect# (дата обращения: 12.03.2018).
9. Качинская И.Б., Сичинава Д.В. О Корпусе диалектных текстов в Национальном корпусе русского языка // Вопросы лексикографии. 2017. № 11. С. 71–85.