РАЗДЕЛ 3. ЯЗЫК — ПОЛИТИКА — КУЛЬТУРА

УДК 81'42
ББК Ш105.51 *ГСНТИ 16.31.21* *Код ВАК 10.02.21*

Ю. В. Богоявленская
Екатеринбург, Россия

РЕПРЕЗЕНТАТИВНОСТЬ ЛИНГВИСТИЧЕСКОГО КОРПУСА: МЕТОД ВЕРИФИКАЦИИ ДОСТОВЕРНОСТИ ПОЛУЧЕННЫХ ДАННЫХ

АННОТАЦИЯ. *Настоящая статья обращена к актуальной проблеме оценки репрезентативности специализированного лингвистического корпуса, предполагающее включение в него необходимо-достаточного количества текстов, обеспечивающих решение исследовательских задач. Анализируется методика достижения репрезентативности корпуса, предложенная А. Н. Барановым. В основе данной методики лежит идея накопления и коррекции относительной частоты феномена, достигаемой в процессе сплошного отбора контекстов его употребления. Метод применим к специализированному корпусу, сформированному исследователем самостоятельно, путем сплошного отбора данных, содержащих исследуемый феномен. Предлагается метод верификации достоверности полученных при индексации корпуса данных, разработанный автором статьи. Метод опирается на закон текстового блока, в соответствии с которым лингвистические единицы (слова, буквы, синтаксические функции, конструкции и т. д.) демонстрируют определенное распределение частоты в одинаково больших текстовых блоках, на принципы итеративности и пропорциональности, а также на принцип вычисления цепного индекса (статистический индекс многошагового расчета, характеризующий изменение показателя по отношению к каждому предыдущему шагу — итерации). На первом этапе рассчитывается индекс достоверности по отдельным параметрам, затем средний индекс достоверности по итерации. На следующем этапе, после новой итерации, средние индексы сопоставляются. Предлагаемый метод обладает такими преимуществами, как обеспечение достижения репрезентативности корпуса, возможность регулировать его объем, а также простота в использовании.*

КЛЮЧЕВЫЕ СЛОВА: *корпус; корпусная лингвистика; репрезентативность; метод верификации достоверности полученных данных.*

СВЕДЕНИЯ ОБ АВТОРЕ: *Богоявленская Юлия Валерьевна, кандидат филологических наук, доцент кафедры романских языков, Уральский государственный педагогический университет; 620017, г. Екатеринбург, пр-т Космонавтов, 26, каб. 465; e-mail : jvbog@yandex.ru.*

В корпусной лингвистике репрезентативность понимается как необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т. п. [Рыков 2002], а корпус рассматривается как некоторый объект, служащий моделью некоторой внешней по отношению к нему реальности. Именно репрезентативность корпуса определяет **достоверность полученных на его материале результатов.** Таким образом, репрезентативность можно рассматривать как проблему адекватного отражения, адаптации или интеграции больших массивов текстов или некоторых иных фрагментов речевой деятельности в существенно меньший по объему корпус текстов [Захаров, Богданова 2011].

Современная корпусная индустрия, стремясь к обеспечению репрезентативности, идет по пути наращивания объема корпусов, исходя из убеждения, что именно ***объем*** и является ключевым моментом для ее достижения. По современным требованиям универсальный (стандартный) корпус должен включать тексты общим объемом не менее 100 млн слов. Утверждается, что достаточно большой (репрезентативный) объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений [Захаров 2005]. Результаты, полученные на материале универсальных корпусов, легко экстраполируются исследователями на весь язык. Однако достаточно ли оснований для подобной экстраполяции?

Собственный опыт корпусного исследования и тщательный анализ научной литературы по данному вопросу убеждает, что в этом вопросе есть слабые места, которые уже начинают осознаваться и акцентироваться в работах, посвященных данной проблеме. Целый ряд исследователей отмечают, что на данный момент *апробированных способов обеспечения репрезентативности корпусов не предложено* [Беликов, др. 2013; McEnery, Hardie 2011; Hunston 2008; Arbach, Ali 2014]. В некоторых работах подчеркивается, что применительно к общеязыковому (универсальному) корпусу это понятие невозможно рассчитать и описать строго математически [Захаров, Богданова 2011]. К тому же языковая личность составителя корпуса оказывает непосредственное влияние на его репрезентативность [Мордовин 2009: 34], что мешает корпусу объективно отражать речевую действительность.

Шаткость категорических заявлений о достигнутой репрезентативности проявляется, в частности, в следующем утверждении авторов Национального корпуса русского языка (НКРЯ), размещенном на официальном сайте: «Национальный корпус имеет две важные особенности. Во-первых, он характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит *по возможности все типы* письменных и устных текстов,

представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус *по возможности пропорционально их доле в языке соответствующего периода*».

К сожалению, заявленные критерии сбалансированности и репрезентативности не обеспечиваются доказательствами (характеристики *по возможности все* и *по возможности пропорционально их доле в языке* вряд ли могут считаться убедительными); на сайте корпуса не представлено ни принципов формирования корпуса, ни методик расчета пропорциональности и т. д. Мы обнаружили только одну работу, в которой анализируется лексическая сбалансированность НКРЯ [Ляшевская, Шаров 2009]. Авторы, создатели частотного словаря на материале данного корпусного ресурса, провели его сегментирование, подсчитали частоты с учетом равномерности распределения по сегментам и пришли к выводу о том, что относительные частоты в ядре языка устроены в НКРЯ близко к тому, что показывает Интернет.

Подобная ситуация наблюдается и в других универсальных корпусах: британском, венгерском, болгарском и т. д. (см. об этом: [Беликов, др. 2013]).

В области **специализированных корпусов**, тексты которых относятся к определенному стилю, сфере коммуникации, дискурсу, автору и т. д., ситуация несколько иная. Требование репрезентативности здесь звучит как *необходимо-достаточное количество текстов, обеспечивающих решение исследовательских задач*. Но как определить, что отобранный исследователем материал действительно достиг этого показателя?

Интересную ***методику достижения репрезентативности корпуса*** предлагает А. Н. Баранов. Под репрезентативностью лингвист понимает «такой тип отображения проблемной области в корпусе данных, при котором последний отражает все свойства проблемной области, релевантные для данного типа лингвистического исследования, в определенной пропорции, определяемой частотой изучаемого явления в проблемной области. Другими словами, относительная частота явления в корпусе должна быть близка его относительной частоте в проблемной области» [Баранов 2001]. Свою стратегию организации корпуса данных он предлагает назвать пропорциональной. Она может применяться к специализированному корпусу, сформированному исследователем самостоятельно, путем сплошного отбора данных, содержащих исследуемый феномен. В основе методики оценки репрезентативности корпуса лежит идея накопления и коррекции относительной частоты феномена, достигаемая в процессе сплошного отбора контекстов его употребления. Автор пишет: «Предположим, что метафорическая модель *M* в проблемной области имеет относительную частоту F_m. Пусть в некоторой выборке примеров A_1 эта модель получает частоту F_1, которая, разумеется, отличается от частоты F_m, но тогда следует ожидать, что при тех же принципах отбора примеров при дальнейшем накоплении материала, равном выборке примеров A_2, в результирующей выборке $A_1 + A_2$ относительная частота F_2 будет в большей степени приближаться к F_m. Конечно, это не обязательно произойдет сразу на втором замере в выборке $A_1 + A_2$. В общем случае это может произойти на n-м замере в выборке $A_1 + A_2 + \ldots + A_n$. С какого-то момента, когда относительная частота метафорической модели начинает приближаться к относительной частоте этой модели в проблемной области, то есть с замера n, различие между относительными частотами F_n, F_{n+1}, F_{n+2} и т. д. будет все больше уменьшаться. Иными словами, разница между числами последовательности Q_n (= F_n — F_{n-1}), Q_{n+1} (= F_{n+1} — F_n), Q_{n+2} = (F_{n+2} — F_{n+1}), характеризующими различия в относительной частоте между замерами, начиная с замера n, будет становиться все меньше и меньше. Графически это должно отражаться движением кривой к некоторому пределу (относительной частоте модели M в проблемной области — F_m)» [Баранов 2001]. Осуществлять замеры и производить расчеты параметров предлагается на материале нескольких метафорических моделей, отобранных из корпуса произвольно (в статье анализируется 5 моделей).

Поставленный А. Н. Барановым эксперимент подтвердил обоснованность подобного подхода. Его недостатком может считаться, на наш взгляд, большая трудоемкость, связанная с ручными расчетами показателей по каждому замеру относительной частоты F, параметров отклонения Q и с исчислениями по разработанным формулам. Однако подход имеет доказанную эффективность, что позволяет использовать его в качестве опоры.

Для проверки достоверности статических данных и оценки репрезентативности корпуса мы предлагаем использовать разработанный нами специальный метод верификации достоверности полученных данных (мы выражаем искреннюю благодарность за консультирование по вопросу разработки мето-

дики зав. кафедрой теоретической механики и математического моделирования УрФУ, кандидату физико-математических наук Михаилу Геннадьевичу Близорукову)**.**

Метод опирается:

1) на закон текстового блока, в соответствии с которым лингвистические единицы (слова, буквы, синтаксические функции, конструкции и т. д.) демонстрируют определенное распределение частоты в одинаково больших текстовых блоках;

2) принципы итеративности (повторяемости, цикличности) и пропорциональности;

3) принцип вычисления *цепного индекса*, широко применяемого в математической, экономической и других статистиках [Гусаров, Кузнецова 2008]. Под цепным индексом понимается *статистический индекс многошагового расчета*, характеризующий изменение показателя по отношению к каждому предыдущему шагу (итерации). Таким образом, базой для сравнения является индекс, полученный в предыдущей итерации.

Мы рассчитываем значение **индекса достоверности по параметру**, используя относительные показатели — индексы параметров, представляющие собой отношение количества употреблений исследуемого феномена, индексированного по данным параметрам, к общему количеству его употреблений.

С целью обеспечения объективности рекомендуется отобрать восемь-десять разноформатных параметров, по которым проводится индексация в корпусе.

После исчисления значения индекса достоверности по каждому параметру в итерации высчитывается **средний индекс достоверности**: сумма индексов по параметрам делится на их количество. После каждой итерации вычисляется значение индекса, его значение сопоставляется с предыдущим до тех пор, пока значение индекса не приблизится к **1.**

Размер итерации определяется индивидуально: это может быть от 100 до 500 (возможно, больше) проиндексированных единиц исследования.

После каждой итерации «снимаются статистические показания», фиксируются в программе «Excel» и вычисляется индекс достоверности каждого параметра по следующей формуле:

$$i_1=\frac{P1}{P0}\ ,\ i_2=\frac{P2}{P1}\ ,\ i_3=\frac{P3}{P2}\ ,\ \dots\ i_n=\frac{Pn}{Pn-1}$$

где

i_1 — значение индекса, получаемое делением относительной частоты параметра, зафиксированной после второй итерации $\mathbf{P_1}$ (***P*** — значение выбранного для исчисления индекса достоверности параметра) на относительную частоту, зафиксированную после первой итерации $\mathbf{P_0}$;

i_2 — значение индекса, получаемое делением относительной частоты параметра, зафиксированной после третьей итерации $\mathbf{P_2}$ на относительную частоту, зафиксированную после второй итерации $\mathbf{P_1}$, и т. д.

Метод был апробирован на специализированном массмедийном сопоставительном корпусе, созданном при помощи программы «Linguistica» (свидетельство о государственной регистрации № 2014660349 от 06.10. 2014). Данное специализированное программное обеспечение предназначено для строительства лингвистических корпусов и позволяет исследователю самостоятельно выстраивать деревья зависимостей (параметрические деревья) и снабжать разметкой как интересующие фрагменты, так и тексты. Программа снабжена системой поиска, статистической обработки результатов; имеется возможность получения конкорданса — генерируемого программой списка фрагментов по заданным параметрам с доступом к источнику. Программа сама просчитывает относительную частотность, поэтому исследователю достаточно сохранить эти показатели в виде таблицы при нескольких итерациях для последующего сопоставления данных.

Для расчетов можно использовать и другие аналогичные программы или воспользоваться программой «Excel».

Применение индекса достоверности позволяет отслеживать текущие изменения статистического процесса. Достоверность может считаться достигнутой при $i_n = 1$, однако такой показатель считается идеальным и практически недостижимым, поэтому мы принимаем допустимыми колебания в определенном промежутке. Данное значение устанавливается индивидуально, но для достижения достоверности важно, чтобы четко была обозначена тенденция приближения к **1.**

Преимуществами предлагаемого метода является обеспечение достижения репрезентативности корпуса, возможность регулировать его объем и простота в использовании.

ЛИТЕРАТУРА

1. Баранов А. Н. Проблема репрезентативности корпуса данных (на примере политической метафорики) // Труды Междунар. семинара «Диалог 2001». — М. : Наука, 2001. URL: http://www.dialog-21.ru/Archive/2001/volume2/2_5.htm.

2. Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П., Шаров С. А. Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и

интеллектуальные технологии. 2013. Вып. 12 (19). Т. 1. С. 84—95.

3. Гусаров В. М., Кузнецова Е. И. Статистика. 2-е изд., перераб. и доп. — М. : ЮНИТИ-ДАНА, 2008. 479 с.

4. Захаров В. П. Корпусная лингвистика. — СПб., 2005. 48с.

5. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. — Иркутск : ИГЛУ, 2011.

6. Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка. — М. : Азбуковник, 2009. URL: http://www.dialog-21.ru/digests/dialog2008/materials/html/53.htm.

7. Мордовин А. Ю. К вопросу о понятии о репрезентативности корпуса текстов // Вестн. ИГЛУ. Сер.: Филология. 2009. № 1. С. 31—37.

8. Рыков В. В. Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Междунар. семинара «Диалог 2002». — М. : Наука, 2002.

9. Arbach N., Ali S. Aspects théoriques et méthodologiques de la représentativité des corpus CORELA // Statut et utilisation des corpus en linguistique. 2014. URL: http://corela.revues.org/3029?lang=en.

10. McEnery T., Hardy A. Corpus linguistics. — Cambridge : Cambridge Univ. Pr., 2011.

Y. V. Bogoyavlenskaya
Ekaterinburg, Russia

REPRESENTATIVENESS OF TEXT CORPUS: METHOD OF VERIFICATION OF DATA RELIABILITY

ABSTRACT. The article addresses the current problem of evaluation of representativeness (reliability) of the specialized linguistic corpus. The requirement to representativeness of a specialized corpus implies the presence of the necessary number of texts that are enough to solve the research problems. The method of representativeness achievement worked out by A.N. Baranov is analyzed. The basis of the method is the idea of accumulation and correction of the relative frequency of the phenomenon achieved by continuous sampling of the contexts it is used in. The method can be applied to a specialized corpus made by the researcher himself by means of continuous sampling of the data with the necessary phenomenon. The method of verification of the received data is offered by the author of this article. The method is based on the law of test block (linguistic units (words, letters, syntactical functions, constructions, etc.) show certain frequency in the equally large text blocks), on the principle of iteration and proportion and on the principle of chain index (statistical index of multi-step calculation that characterizes the change of index compared to the previous step — iteration). On the first stage we calculate the index of reliability on certain parameters, then the index of reliability in iteration. On the next stage, after new iteration, the average indices are compared. The advantages of this method are representativeness of the corpus, the possibility of regulating the amount of data in the corpus and it is easy to work with.

KEYWORDS: corpus; corpus linguistics; representativeness; method of verification of reliability of data.

ABOUT THE AUTHOR: *Bogoyavlenskaya Yulia Valerievna, Candidate of Philology, Associate Professor of Department of Romance Languages, Ural State Pedagogical University, Ekaterinburg, Russia.*

REFERENCES

1. Baranov A. N. Problema reprezentativnosti korpusa dannykh (na primere politicheskoy metaforiki) // Trudy Mezhdunar. seminara «Dialog 2001». — M. : Nauka, 2001. URL: http://www.dialog-21.ru/Archive/2001/volume2/2_5.htm.

2. Belikov V. I., Kopylov N. Yu., Piperski A. Ch., Selegey V. P., Sharov S. A. Korpus kak yazyk: ot masshtabiruemosti k differentsial'noy polnote // Komp'yuternaya lingvistika i intellektual'nye tekhnologii. 2013. Vyp. 12 (19). T. 1. S. 84—95.

3. Gusarov V. M., Kuznetsova E. I. Statistika. 2-e izd., pererab. i dop. — M. : YuNITI-DANA, 2008. 479 s.

4. Zakharov V. P. Korpusnaya lingvistika. — SPb., 2005. 48s.

5. Zakharov V. P., Bogdanova S. Yu. Korpusnaya lingvistika. — Irkutsk : IGLU, 2011.

6. Lyashevskaya O. N., Sharov S. A. Chastotnyy slovar' sovremennogo russkogo yazyka. — M. : Azbukovnik, 2009. URL: http://www.dialog-21.ru/digests/dialog2008/materials/html/53.htm.

7. Mordovin A. Yu. K voprosu o ponyatii o reprezentativnosti korpusa tekstov // Vestn. IGLU. Ser.: Filologiya. 2009. № 1. S. 31—37.

8. Rykov V. V. Korpus tekstov kak realizatsiya ob"ektno-orientirovannoy paradigmy // Trudy Mezhdunar. seminara «Dialog 2002». — M. : Nauka, 2002.

9. Arbach N., Ali S. Aspects théoriques et méthodologiques de la représentativité des corpus CORELA // Statut et utilisation des corpus en linguistique. 2014. URL: http://corela.revues.org/3029?lang=en.

10. McEnery T., Hardy A. Corpus linguistics. — Cambridge : Cambridge Univ. Pr., 2011.

Статью рекомендует к публикации д-р филол. наук, проф. А. П. Чудинов.