УДК 81

КОРПУСНОЕ ИССЛЕДОВАНИЕ ЯЗЫКА: МОДА ИЛИ НЕОБХОДИМОСТЬ?

О. О. Борискина

Воронежский государственный университет

Поступила в редакцию 20 апреля 2015 г.

Аннотация: *в статье обсуждается роль и место корпусных исследований языка в современной науке, сопоставляются аргументы за и против применения корпусных технологий в лингвистическом описании. Выявляются основания, по которым возможно судить о значимости и перспективности корпусных исследований.*
Ключевые слова: *корпусное исследование, количественный анализ, узус, интроспекция.*

Abstract: *this paper discusses the role of corpus-based and corpus-driven study of a language in modern linguistics. Arguments for and against corpus technology are provided. The paper is also concerned with the grounds how the significance of corpus research and its perspectives can be evaluated.*
Key words: *corpus-based study, corpus-driven study, quantitative analysis, usage, introspection.*

Вопрос роли и места корпусного исследования языка в современной науке отнюдь не праздный, если до сих пор на обсуждениях диссертаций и заседаниях диссертационных советов использование корпусных данных в исследовании языковых явлений встречает «холодный прием», сопровождающийся сомнениями в целесообразности применения корпусных методов в лингвистическом описании и скептическими отзывами о надежности такого источника информации, как корпус.

Так что же такое корпусное исследование языка: требование времени, сформировавшееся как новое направление в лингвистике, или модная тенденция, временно господствующая в языковых штудиях? В современной науке существуют разные точки зрения насчет оценки функциональности и значимости корпусного исследования языка. Ряд исследователей признают за ним статус основной эмпирической лингвистической парадигмы, другие предпочитают пользоваться корпусом исключительно как источником примеров для иллюстрации положений своих теорий.

Крайне радикальным представляется полное неприятие корпусной лингвистики (ср. высказывание Н. Хомского в интервью 2004 г. «Corpus linguistics doesn't mean anything»). Согласно логике отца генеративной лингвистики корпусный подход сводится к простому наблюдению за большим объемом данных, что «не является методом научного познания и не может обеспечить ни успешного решения познавательных и практических проблем, ни приращение знания» [1].

Основными оппонентами генеративистов по вопросам познавательного потенциала корпусных исследований являются представители корпусной лингвистики.

Авторы учебного пособия по корпусной лингвистике [2] представляют всю палитру современных корпусных исследований, показывая, что различия между двумя основными подходами к изучению языковых явлений *Corpus-driven* vs *Corpus-based* стираются. Если в исследовании, основанном на данных корпуса (*corpus-based study*), решается вопрос проверки валидности теории или гипотезы с использованием корпусных методов, то лингвист-корпусник, «*движимый данными*» корпуса (*corpus-driven study*), строит свою теорию, полностью и всецело полагаясь на материал корпуса, описывая таким образом узус. Как видим, в мировой науке не стоит вопрос об обращении к корпусам, вопрос заключается скорее в подходе к этому лингвистическому ресурсу, да и это отличие весьма условно. Представители когнитивного направления в зарубежном языкознании инкорпорируют эмпирические методы в лингвистическое описание, приравнивая в правах эксперимент и корпусные данные.

В среде отечественных лингвистов также нет единодушия в отношении к корпусам. Большинство исследований, опирающихся на корпусные данные, проводится в рамках массового направления «корпус как инструмент», что во многом напоминает *corpus-based* подход к лингвистическому описанию, в то время как создатели Национального корпуса русского языка руководствуются принципом «корпус как идеология», ориентируясь на методологию и терми-

нологию *corpus-driven* подхода. По мнению романтически настроенной части лингвистического сообщества, «в лингвистике произошла корпусная революция. После появления корпусов эта наука стала совсем другая»; и даже если «этот пафос немного убрать, чуть-чуть снизить градус, то степень значимости все-таки останется. Корпусное исследование – это больше, чем методика анализа», – считает В. А. Плунгян [3]. Это направление представлено лингвистами, не только использующими корпусные методы или данные в своей работе, но и создающими и аннотирующими корпусные ресурсы (см., например, НКРЯ [4]). Накопленный в процессе такой работы опыт питает их убежденность в том, что «современная лингвистика должна стать лингвистикой корпусов» [3], когда отбор материала исследования будет основываться не на данных словарей и других лексикографических источников, а будет осуществляться при помощи грамотно сформированных поисковых запросов.

Представители сдержанно-скептического, осторожного отношения к использованию корпусных технологий в лингвистике высказывают мнение, что обращение к корпусным ресурсам в лингвистическом описании – это отчасти дань моде. Как на рубеже веков модно было изучать концепты, скрипты, фреймы и другие заимствованные из западной лингвистики конструкты, так и сегодня считается модным привлекать корпусные данные, применять количественный и статистический анализ, устанавливать частотность использования языковых единиц и изучать их встречаемость в конструкциях и коллострукциях. Однако, убеждены скептики, увлечение корпусами пройдет, как прошла мода на описание концептов.

Стоит упомянуть и попытки российских когнитологов интегрировать корпусные, в том числе и количественные и статистические данные в ставшие традиционными для отечественной лингвистики XXI в. когнитивные исследования. Российская когнитивная лингвистика ищет пути интеграции с корпусной с целью «разработать интегрированную методологию в двух современных парадигмах когнитивной и корпусной» [5]. С результатами синтеза двух направлений можно познакомиться в недавно вышедшей коллективной монографии «Методы когнитивного анализа семантики слова», где анализ семантики языковых единиц проводится с опорой на корпусные данные или с привлечением корпусных данных [5].

Рассмотрим доводы, которые приводят сторонники «безкорпусного» лингвистического анализа для аргументации своей позиции.

Казалось бы, интенсивное развитие информационных технологий не дает повода для сомнений в необходимости и перспективности корпусной лингвистики и корпусного исследования языка. Однако скептики убеждены, что нелепо говорить о перспективности направления, у которого нет даже собственного УДК-классификатора (приводимый часто УДК-81’32 – это «Математическая лингвистика»). И этот формальный признак далеко не единственный, дискредитирующий значимость корпусного исследования языка.

Другим аргументом, распространенным в среде скептически настроенной части лингвистического сообщества, является сомнение в целесообразности наделять исследование по данным корпуса особым статусом, поскольку в любой области языкознания лингвист формирует материал в виде картотеки. Это обязательный этап любого научного поиска, будь то лингвистический эксперимент или полевая работа. *Корпус*, по их мнению, – модное слово, вытесняющее из русского языка родное – *картотека*. Очевидно, здесь мы наблюдаем подмену понятий. Корпус (с разметкой и аннотированием) – это далеко не картотека и по масштабу, и по функциональности, и по возможностям.

Ю. Д. Апресян, один из сторонников взвешенного, осторожного обращения к корпусным данным, особенно при обращении к Wikipedia Corpus или Google Books, считает, что некоторые лингвисты слепо следуют «моде на корпус», и это повальное увлечение зачастую ведет к фальсификации результатов и злоупотреблению количественными данными. Сырой частотный подсчет употребления слов не может выступать критерием истинности утверждений о функционировании лингвистического объекта. Данный вопрос качества полученных результатов всецело связан с профессиональной компетенцией исследователя. Проявлениям дилетантства и кустарничества в среде лингвистов-корпусников отчасти «способствуют» и неустоявшийся терминологический аппарат и неотлаженная методология корпусного исследования, но данное обстоятельство никак не умаляет значимости корпусных технологий.

Связанный с этим следующий аргумент, свидетельствующий о «моде на корпус», состоит в том, что многие электронные лингвистические ресурсы – платные. В такой ситуации российские ученые, которые хотят «быть в тренде», используют демоверсию, что зачастую умалчивается в научных трудах. Такая практика приводит к появлению откровенной халтуры, поскольку результаты анализа основаны на неточных и неполных данных демоверсий.

Еще одним аргументом в пользу ненадежности получаемого в рамках корпусного исследования результата является узуальная природа корпуса, содержащего ненормативное и неправильное употребление, в частности это касается интернет-ресурсов. Такое положение вещей оправдывает опасения сторонников традиционных методов исследования и

интроспекции в ограниченности корпусного подхода. Фактически, результаты такого исследования ***ограничены*** описанием узуса, что, по мнению скептиков, не позволяет делать выводы о теоретически значимых закономерностях языковой системы.

Применение корпусных технологий предполагает знакомство с основами корпусной и IT-терминологии, владение навыками формирования оптимального для целей исследования поискового запроса и методами количественной и статистической обработки данных. Затрудняет работу лингвиста-корпусника и несовершенство поискового инструментария, что порождает определенную долю «шума». Поиск по запросу может выдавать сотни и даже тысячи результатов (контекстов словоупотребления), которые просто физически нереально просмотреть в ограниченное время. Это провоцирует «скептиков» критически относиться к заявлениям «революционеров» и «романтиков» о том, что корпусные технологии экономят время, а поисковая система способствует решению проблем устройства и развития языка. Усовершенствование поисковиков и методик запроса – одна из важных задач, стоящих перед корпусной лингвистикой, решению которой много внимания уделяется в рамках международной научной конференции «Корпусная лингвистика» (http://mathlingvo.ru/2014/10) и Международной научной конференции «Диалог» (http://www.dialog-21.ru/).

Теперь рассмотрим, на чем основана убежденность лингвистов-корпусников в том, что современное исследование языка не может быть проведено вне «лингвистики корпусов». Во-первых, корпусные исследования языка отличаются бóльшей представительностью данных, что предполагает и реально квантитативные, и статистические исследования (см., например: [6; 7]). Однако вопрос о сбалансированности и репрезентативности корпуса с повестки дня не снимается [8]. Во-вторых, корпусные исследования – это все более и более удачные попытки из необозримого (например, устного дискурса) сделать обозримое (дискурс, представленный в размеченных текстах, который можно изучать [9]). Кроме того, корпусные технологии позволяют приступить к наблюдению за редкими языковыми явлениями и проследить динамику языковых изменений на малом временном отрезке. В корпусном исследовании языка находят отражение и получают интерпретацию как частотные явления, так и окказиональные. Решение некоторых исследовательских задач предполагает обращение не к одному, а нескольким ресурсам. Сопоставляя и грамотно анализируя данные, полученные с помощью различных корпусов, возможно установление языковой вариативности и закономерностей языковых изменений, предсказание дальнейшего развития описываемого явления, а также осмысление таких употреблений, которые противоречат установившимся представлениям о языковой норме.

В отличие от других видов исследования (интроспекции, словарной или полевой работе), корпусное позволяет проверять гипотезы о языковых изменениях и закономерностях. «…Использование корпусов делает возможным объективизировать лингвистику, найти более веские аргументы применительно к фактам, создать ситуацию повторяемости, что, по мнению А. Мустайоки, является важнейшим элементом науки» [4]. Именно проверяемость результатов корпусного исследования обеспечивает его эффективность, достоверность и повторяемость. Корпусный подход к исследованию делает результаты более эмпирически релевантными. «Лингвистика корпусов позволяет нам понять, каков язык на самом деле, а не каким мы хотим, чтоб он был», – считает В. А. Плунгян [3].

Еще одним фактором, свидетельствующим о необходимости корпусного исследования, является тот факт, что корпусные технологии позволяют получить принципиально новые данные о том, как устроен язык и как он функционирует.

В. А. Плунгян предостерегает об «опасности новизны» в лингвистике, называя это парадоксом внутреннего развития лингвистики. «Можем мы узнать многое, мы, лингвисты, а вот хотим ли мы это знать? Оказывается, что не всегда и не все лингвисты этого хотели, это очень интересный факт. Огромный массив данных, которые буквально хлынули на нас, во многом может заставить пересмотреть существующие представления о языке, о том, что это такое, как он существует, как он изменяется. Понятно, что это не всем может понравиться, у всех представлений могут быть авторы, эти авторы как-то существуют в науке, а тут появляется вдруг какой-то корпус, из которого следует, что всё не так, что нужны новые идеи, новые теории. Лучше уж мы будем как раньше. Психологически это вполне понятно» [3], но неприемлемо и с точки зрения практической значимости корпусных исследований.

Принципиальная новизна результатов исследования позволяет говорить о правомерности создания «корпусных словарей» и «корпусных грамматик» нового поколения, выполненных – и верифицированных – именно по отношению к конкретному фиксированному корпусу. Корпусный характер словарей и грамматики повышает их надежность и проверяемость, позволяет избегать той субъективности и неполноты, которыми часто страдают описания, опирающиеся исключительно на интроспекцию лингвиста. Создание анализаторов и специализированных словарей для автоматизированной настройки разметки корпуса (морфологической, синтаксической, тематической или семантической) технологически возмож-

но только в рамках лингвистики корпусов. Другим практическим достижением корпусных технологий является существенное уменьшение трудоемкости процесса сбора и обработки материала (в человеко/часах). Для получения тех же данных вручную (например, путем простого просмотра текстов и выписывания примеров на карточки, как это происходило в докомпьютерную эпоху) могут потребоваться месяцы и даже годы. Это, увы, не относится к созданию корпусных ресурсов, а только к их использованию.

Знакомство с результатами отечественных и зарубежных корпусных исследований языка, оценка их теоретической и практической значимости, их принципиальной новизны и перспективности являются достаточным основанием для определения роли и места корпусного исследования языка в современной науке. Взвесив сильные и слабые стороны корпусного исследования языка, нельзя не заметить, что корпус является средой для получения новых научных данных, осмысление которых представляется приоритетным для современного лингвистического описания и абсолютно необходимым в научной деятельности современного исследователя. Принимая во внимание доводы оппонентов и опасения скептиков по поводу необходимости и целесообразности корпусных исследований, различая в корпусном исследовании признаки модного увлечения, следует признать, что это – требование времени, связанное с новым качеством лингвистической реальности и отвечающее потребностям современного общества.

ЛИТЕРАТУРА

1. *Andor J.* The master and his performance : An interview with Noam Chomsky // Intercultural Pragmatics 1-1 (2004), 93–111.

2. *McEnery T., Hardie A.* Corpus Linguistics : Method, theory and practice. Cambridge : Cambridge University Press. 2012. Support website for Corpus Linguistics : Method, theory and practice. – Mode of access : http://corpora.lancs.ac.uk/clmtp

3. *Плунгян В. А.* Почему современная лингвистика должна быть лингвистикой корпусов. Лекция, прочитанная в рамках проекта «Публичные лекции». – Режим доступа: http://polit.ru/article/2009/10/23/corpus/

4. Национальный корпус русского языка. – Режим доступа: www.ruscorpora.ru

5. Методы когнитивного анализа семантики слова : компьтерно-корпусный подход / под общ. ред. В. И. Заботкиной. – М. : Языки славянской культуры, 2015. – 344 с.

6. *Boriskina O. O.* A Corpus-based Study of Noun Cryptotypes in English / O. O. Boriskina // Компьютерная лингвистика и интеллектуальные технологии : материалы ежегодной Междунар. конф. / ред. кол. : А. Е. Кибрик (гл. ред.), В. И. Беликов, И. М. Богуславский, Б. В. Добров, Д. О. Добровольский [и др.]. – 2011. – С. 135–145.

7. *Донина О. В.* Криптоклассные данные для определения меры языковой эквивалентности / О. В. Донина // Вестник Воронеж. гос. ун-та. Сер. : Лингвистика и межкультурная коммуникация. – 2015. – № 1. – С. 108–110.

8. *Шилихина К. М.* Роль контекста в интерпретации иронии / К. М. Шилихина // Вестник Воронеж. гос. ун-та. Сер. : Лингвистика и межкультурная коммуникация. – 2008. – № 2. – С. 10–15.

9. *Шилихина К. М.* Использование корпусов в исследованиях дискурса / К. М. Шилихина // Вестник Воронеж. гос. ун-та. Сер. : Лингвистика и межкультурная коммуникация. – 2014. – № 3. – С. 21–26.

Воронежский государственный университет

Борискина О. О., доктор филологических наук, доцент кафедры английского языка в профессиональной международной деятельности

E-mail: olboriskina@gmail.com

Voronezh State University

Boriskina O. O., Doctor of Philology, Associate Professor of the English for International Relations Department

E-mail: olboriskina@gmail.com