УДК 81

ГЕНЕРАЛЬНЫЙ ИНТЕРНЕТ-КОРПУС РУССКОГО ЯЗЫКА И ПОНЯТИЕ РЕПРЕЗЕНТАТИВНОСТИ В КОРПУСНОЙ ЛИНГВИСТИКЕ

Пиперски А.Ч.[1]

[1]*Институт лингвистики ФГБОУ ВПО «Российский государственный гуманитарный университет», Москва, Россия (125993, г. Москва, Миусская пл., 6, корп. 2), e-mail: apiperski@gmail.com*

В данной статье анализируется использование понятия репрезентативности в корпусной лингвистике и делается вывод о том, что в отсутствие точных методов оценки репрезентативность корпуса определяется негласной договоренностью между создателями корпуса и его пользователями. Разрабатываемый в настоящее время Генеральный интернет-корпус русского языка (ГИКРЯ) задумывается как инструмент, позволяющий эксплицировать подобные договоренности и изучать русский язык в его дифференциальной полноте. Исследователи получат ресурс, позволяющий анализировать отдельные сегменты Интернета и создавать подкорпуса на основе метаразметки, извлекаемой автоматически. В настоящее время в ГИКРЯ размечены и доступны для поиска два сегмента русского Интернета: блог-платформа LiveJournal.com и «Журнальный зал». В дальнейшем количество сегментов планируется существенно расширить.

Ключевые слова: корпусная лингвистика, Интернет, репрезентативность, метаразметка.

THE GENERAL INTERNET CORPUS OF RUSSIAN AND THE NOTION OF REPRESENTATIVENESS IN CORPUS LINGUISTICS

Piperski A.C.[1]

[1] *Russian State University for the Humanities, Institute of Linguistics, Moscow, Russia (Miusskaya Sq. 6-2, 125993, Moscow), e-mail: apiperski@gmail.com*

The present article deals with the notion of representativeness in corpus linguistics. It turns out that there are no exact methods for assessing representativeness, and for this reason the representativeness of a corpus is nothing more than a tacit agreement between the creators of a corpus and its users. The General Internet Corpus of Russian (GICR) which is presently under development tries to make such an agreement explicit. It encourages its uses to study register variation in the Russian language of the Internet. The linguistic community will be able to use a research tool to study different segments of the Web and to create subcorpora using automatically extracted metadata. As for June 2013, GICR contains two segments of the Russian Web, namely the blog platform LiveJournal.com and the "Magazine Reading Room" (http://magazines.russ.ru/). More segments will be added soon.

Keywords: corpus linguistics, Internet, representativeness, metadata.

Корпусные исследования являются одним из важнейших направлений современной лингвистики, которое позволяет получать объективные данные о языке, не прибегая к интроспекции, которая часто влечет за собой необъективность. Корпусами называются информационно-справочные системы, основанные на собрании текстов на некотором языке (или языках) в электронной форме и снабженные разметкой [4]. Разработка корпусов — это сложная задача, в решении которой лингвисты должны сотрудничать с инженерами. Кроме того, именно корпусные исследования заставляют лингвистов вводить в свой научный обиход некоторые новые понятия. Например, вопрос о репрезентативности выборок, чрезвычайно релевантный для многих других наук, раньше не был особенно актуален для

лингвистов, но теперь, в связи с распространением корпусов, он получает первостепенное значение [6].

Лингвисты стремятся делать свои выводы максимально общими: мало кому хочется говорить об особенностях собственного идиолекта (т.е. своего индивидуального языка) или об особенностях идиолекта информанта. Предпочтительными являются суждения о том или ином языке в целом (например, о русском языке, об английском языке, о языке бурушаски) или об отдельных, но достаточно крупных частях языка (о русской разговорной речи, об английском языке Канады и т.п.). Однако для того, чтобы иметь возможность говорить о языке в целом или о той или иной его части, необходимо иметь репрезентативный источник данных.

Репрезентативность корпусов — это не такой простой вопрос, как может показаться на первый взгляд. Очевидно, что никакой корпус не может включать в себя всех текстов на том или ином языке, а значит, тексты, входящие в корпус, неизбежно представляют собой некоторую выборку. Вопрос о том, можно ли считать данные, полученные на этой выборке, масштабируемыми на весь язык, часто приходится решать самому исследователю.

Если исследователь может найти в корпусе примеры на интересующее его явление, то его дальнейшее поведение во многом зависит от того, как позиционируется этот корпус. В названии некоторых корпусов содержится информация об их составе (таков, например, Мичиганский корпус устного академического английского языка – Michigan Corpus of Academic Spoken English [7]), и пользователи таких ресурсов едва ли рискнут масштабировать полученные результаты на весь язык. Однако корпуса, позиционирующие себя как национальные, намного сильнее навязывают своим пользователям представление, что по ним можно делать выводы про язык в целом.

Обычно корпуса пытаются так или иначе обосновать, что они хорошо представляют язык в целом. Масштабируемость выводов на язык связана с понятиями сбалансированности и представительности, которые часто прямо или косвенно упоминаются в описаниях корпусов:

Болгарский национальный корпус постоянно развивается и пополняется новыми текстами, ставя перед собой цель достичь представительности и сбалансированности благодаря включению текстов разных способов бытования (письменных и устных), разных эпох и разнообразной тематической и жанровой принадлежности.[1] [3]

[1] Българският национален корпус постоянно се развива и обогатява с нови текстове, като се цели постигането на представителност и балансираност чрез включването на текстове от различна модалност (писмени и устни), различни периоди на създаване, разнообразни тематични области и жанрове.

Национальный корпус [русского языка] 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 [4].

Какими свойствами обладает Британский национальный корпус?

<…>

Генеральность: он [Британский национальный корпус] включает в себя много различных стилей и разновидностей языка, не ограничиваясь какой-либо определенной тематической областью, жанром или регистром. В частности, в нем содержится как устный, так и письменный язык [2] [5].

Подобные утверждения чаще всего приводятся без особых доказательств, и пользователям остается принимать их на веру. Фактически получается, что репрезентативность корпуса — это результат негласного договора между его создателями и пользователями [2].

Задача разрабатываемого в настоящий момент Генерального интернет-корпуса русского языка (ГИКРЯ) [1; 2] заключается в том, чтобы сделать этот договор между создателями корпуса и лингвистами эксплицитным. Пользователи корпуса должны отдавать себе отчет в том, как устроен корпус и какие тексты входят в те или иные его части.

В основе ГИКРЯ лежит понятие сегмента Интернета: производится как можно более полная выкачка тех или иных частей сети, которые кажутся интересными с лингвистической точки зрения. По состоянию на июнь 2013 года скачаны, полностью проиндексированы, размечены и доступны для поиска два сегмента русскоязычного Интернета — блог-платформа LiveJournal и «Журнальй зал» (http://magazines.russ.ru). Исследования, выполненные на основании этих данных, могут быть достаточно надежно масштабированы на русский язык блогов и на русский язык современной художественной литературы и публицистики. Следует отметить, что создатели ГИКРЯ призывают не обобщать результаты на русский язык в целом, а внимательно относиться к тому, на материале каких текстов они были получены

Поскольку ГИКРЯ стремится к дифференциальной полноте, состав сегментов Интернета, входящих в корпус, в будущем будет существенно расширен. ГИКРЯ

---

[2] What sort of corpus is the BNC?

<…>

General: It includes many different styles and varieties, and is not limited to any particular subject field, genre or register. In particular, it contains examples of both spoken and written language.

основывается на автоматических методах скачивания и разметки, и поэтому увеличение размера корпуса не будет сопряжено с увеличением объемов ручной работы.

Впрочем, пользователи смогут осуществлять поиск не только по априорно выделенным сегментам Интернета, но и по текстам, отобранным по другим признакам. В частности, LiveJournal.com и «Журнальный зал» были выбраны в качестве пилотных сегментов ГИКРЯ именно потому, что из них можно извлечь большое количество метаинформации (пол, возраст, региональная принадлежность говорящего и т.п.), которая также включается в ГИКРЯ. На основании метаразметки строятся классификаторы, которые с высокой долей вероятности приписывают метатекстовые признаки другим текстам, не снабженным подобного рода информацией. Благодаря этому ГИКРЯ можно использовать не только для того, чтобы анализировать язык сегментов Интернета, но и для более традиционных социолингвистических исследований — например, можно изучать различия в языке различных регионов, а также возрастные и гендерные различия.

Примером изучения вариативности в русском языке может служить исследование региональных разновидностей русского языка. В основу регионально размеченного подкорпуса ГИКРЯ легли блоги с платформы LiveJournal.com, поскольку многие пользователи этого ресурса указывают информацию о регионе проживания и получения образования в своем профиле. В настоящий момент выделяется 16 региональных подкорпусов для тех регионов русскоговорящих стран, для которых удалось собрать достаточно большое количество данных (таблица 1).

Таблица 1 – Перечень региональных подкорпусов.

Регион	**Страна**	**Кол-во документов**	**%**
Донецкая область	Украина	39 080	3,14%
Киев	Украина	114 736	9,21%
Краснодарский край	Россия	50 544	4,06%
Красноярский край	Россия	41 032	3,29%
Московская область	Россия	119 328	9,58%
Новосибирская область	Россия	78 106	6,27%
Омская область	Россия	32 396	2,60%
Пермский край	Россия	55 226	4,43%
Республика Башкортостан	Россия	53 420	4,29%
Республика Татарстан	Россия	34 684	2,78%
Ростовская область	Россия	64 340	5,17%
Самарская область	Россия	82 450	6,62%

Санкт-Петербург	Россия	300 814	24,15%
Саратовская область	Россия	31 706	2,55%
Свердловская область	Россия	97 894	7,86%
Челябинская область	Россия	49 798	4,00%
	Всего:	1 245 554	100%

Наличие подобного корпуса позволяет исследователям анализировать распределение регионализмов в лексике и грамматике не только интуитивно, но и с помощью надежных статистических данных.

В качестве примера сравним частотность сочетаний *в Украину* и *на Украину* в различных русскоговорящих регионах. На рисунке 1 изображен интерфейс корпуса, позволяющий увидеть частотность употреблений с разбивкой по региональным подкорпусам. Приводится как абсолютное количество вхождений, так и стандартизированный показатель ipm (instances per million, вхождений на миллион).

	в Украину		на Украину	
Корпус	Экземпляров	IPM	Экземпляров	IPM
BASHKIRIYA	10	1.029	13	1.338
PETERSBURG	29	0.639	61	1.345
SVERDLOVSKAYA	11	0.672	26	1.589
KRASNODARSKIY	12	1.403	15	1.754
KRASNOYARSKIY	10	1.230	23	2.829
NOVOSIBIRSKAYA	5	0.379	22	1.669
OMSKAYA	4	0.799	4	0.799
PERMSKIY	5	0.504	21	2.115
ROSTOVSKAYA	9	0.885	16	1.573
SAMARSKAYA	7	0.487	26	1.809
SARATOVSKAYA	1	0.186	6	1.118
TATARSTAN	3	0.455	6	0.910
KIEV	356	16.111	155	7.015
DONETSKAYA	54	4.832	98	8.769
CHELYABINSKAYA	9	1.087	8	0.966
Totals	525	2.702	500	2.573

Рисунок 1 – Интерфейс корпуса, отображающий частотность употреблений с разбивкой по региональным подкорпусам.

На этом материале можно сделать тривиальный вывод о том, что словосочетания *в Украину* / *на Украину* чаще встречаются в текстах, написанных жителями этой страны. Однако гораздо более показательно, что заметное преобладание предлога *в* над предлогом *на* отличает не Украину от России, а только один из представленных в таблице регионов, а именно Киев, от всех остальных, в том числе и от Донецкой области. Следовательно, при

помощи ГИКРЯ исследователь может уловить достаточно тонкое языковое различие между регионами и в дальнейшем может попытаться скоррелировать подобное различие с теми или иными экстралингвистическими факторами (например, с политической обстановкой в различных частях Украины).

Таким образом, ГИКРЯ представляет собой инструмент для исследования русского языка в его разнообразии. Создатели ГИКРЯ не имплицируют существование единого русского языка, а, напротив, стимулируют пользователей корпуса исследовать различные разновидности языка. При этом различия могут проявляться в разных плоскостях: можно исследовать различные сегменты Интернета, различные региональные варианты русского языка, гендерно обусловленные варианты русского языка и так далее.

Работы проводятся при финансовой поддержке Министерства образования и науки Российской Федерации в рамках выполнения ГК 07.514.11.4142 по теме «Разработка методов автоматического и полуавтоматического создания корпуса и подкорпусов современного русского языка на основе русскоязычного Интернета» и программы стратегического развития РГГУ.

Список литературы

1. Беликов В.И., Селегей В.П., Шаров С.А. 2012. Пролегомены к проекту Генерального интернет-корпуса русского языка // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). Вып. 11 (18). - М. : Изд-во РГГУ, 2012. - С. 37–50.

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 : по материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая – 2 июня 2013 г.). Вып. 12 (19). — М. : Изд-во РГГУ, 2013. - С. 84–95.

3. Български национален корпус — описание [Электронный ресурс]. — (http://www.ibl.bas.bg/BGNC_classific_bg.htm) (дата обращения: 01.06.2013).

4. Что такое корпус? [Электронный ресурс]. — (http://ruscorpora.ru/corpora-intro.html) (дата обращения: 01.06.2013).

5. About the British National Corpus [Электронный ресурс]. — (http://www.natcorp.ox.ac.uk/corpus/index.xml) (дата обращения: 01.06.2013).

6. McEnery T., Hardie A. Corpus Linguistics. - Cambridge: Cambridge University Press, 2011. xv, 294 p.

7. Michigan Corpus of Academic Spoken English [Электронный ресурс]. — (http://quod.lib.umich.edu/m/micase/) (дата обращения: 01.06.2013).

Рецензенты:

Беликов В.И., д.ф.н., доцент, кафедра теоретической и прикладной лингвистики филологического факультета Московского государственного университета имени М.В. Ломоносова, г. Москва.

Гриненко М.М., д.ф.-м.н., научный консультант, ООО «Аби ИнфоПоиск», г. Москва.