

Перепечкина С.Е.

УДК 811.112.2:81'33

**КОРПУСНЫЙ ПОДХОД В СОВРЕМЕННЫХ ИССЛЕДОВАНИЯХ
ЯЗЫКОВОЙ ВАРИАТИВНОСТИ**

Аннотация. В статье рассматриваются возможности корпусного подхода в исследовании языка, функционирующего в реальных условиях; указываются характеристики метода, определяющие его надежность и достоверность. Основное внимание уделяется новейшим исследованиям социально и территориально обусловленной вариативности немецкого языка в Северной Германии и стилистической языковой дифференциации в виртуальных жанрах коммуникации; описываются возможности электронных собраний текстов для использования их в научных целях. **Ключевые слова:** лингвистический корпус, корпусный подход, исследование вариативности языка, SiN-проект, чат-корпус.

Анотація. У статті розглядаються можливості корпусного підходу у дослідженні мови, що функціонує в реальних умовах; указуються характеристики методу, які визначають його надійність і достовірність. Основна увага приділяється новітнім дослідженням соціально і територіально зумовленої варіативності мови у Північній Німеччині та стилістичної мовної диференціації у віртуальних жанрах комунікації; описуються можливості електронних колекцій текстів для використання їх з науковою метою.

Ключові слова: лінгвістичний корпус, корпусний підхід, дослідження варіативності мови, SiN-проект, чат-корпус.

Summary. This article discusses the possibilities of corpus-based approach in the linguistic study, i.e. the study of language as expressed in samples (corpora) of "real world" text. It introduces the basic aspects of this approach – corpus representativeness and balance. Being "collections of naturally occurring language text" (John Sinclair), corpora can provide a good basis for the study of language variation with the application of new electronic technologies. The article focuses on the latest research of social and geographical variability of language in Northern Germany as well as on the language differentiation in the virtual genres of communication. With the fact a new spoken language variety, regarded as a contact variety (a result of High and Low German contact), has become widespread in the Northern of Germany, a new research of language situation in this area became necessary. The identification, description and analysis of language variation in terms of regional, social and situational usage are the primary purpose of the SiN project (Sprachvariation in Norddeutschland). Documenting object language data along with the speakers' subjective opinions and reflexions on language norm will allow for compiling the first ever text corpus of Low German substandard dialect. On the bases of such a corpus a wide variety of socio- and idiolect research on phonetic, lexical and grammatical level can be realized. Another actual problem today is the Internet-based communication and its genres: e-mail, forums, web-chats and weblogs, Skype etc., so it becomes necessary to create balanced, annotated corpora of computer-mediated communication such as the Dortmund Chat Corpus (ca. 500 annotated chat-logs and a retrieval tool available online), whose resources can be used for studying of chat-style in its various forms.

Keywords: linguistic corpus, corpus-based approach, linguistic research, variability of language, SiN project, chat corpus.

Описание языка, основанное на корпусном подходе, восходит еще к 13 веку (о нецифровых корпусах, напр., см. [1, с. 13-18; 2, с. 1-14]), однако в своей современной форме существует лишь с начала 1960-х годов [3, с. 15] и соотносится сегодня почти исключительно с наличием электронных собраний данных. основополагающим принципом корпусной лингвистики является использование аутентичного материала, на основе которого можно сделать достоверные выводы о структуре языка и особенностях его функционирования, в т.ч. обусловленных факторами экстралингвистического характера, описать феномены языкового изменения в динамике – проблемы, которые всегда актуальны в науке о языке. Поскольку корпуса представляют собой "collection[s] of naturally occurring language text" [4, с. 171], т.е. состояются из речевых высказываний, возникших в естественном контексте, они могут послужить хорошей базой для изучения языковой вариативности с применением новых технических возможностей.

Цель данной статьи заключается в анализе ряда современных исследований языковых процессов, результатом которых стало создание объемных корпусов текстов, которые, в свою очередь, могут быть использованы в качестве инструмента для дальнейших разработок в лингвистике и решения прикладных задач (напр., в лингводидактике).

По определению Мак Энери, "a corpus is a collection of machine-readable authentic texts [...] which is sampled to be representative of a particular language or language variety" [5, с. 5], из чего следует, что непременным условием для всестороннего исследования языка является аутентичность и репрезентативность текстов, а электронная форма обеспечивает их доступность. Тщательно составленный корпус с максимальной объективностью представляет разнообразие изучаемого явления и дает такую же объективную картину применения этого явления в речевой практике носителей языка.

Одним из грандиозных исследовательских проектов, основанных на корпусном подходе, является сегодня проект под названием „Sprachvariation in Norddeutschland“ / SiN (<http://sin.sign-lang.uni-hamburg.de>), рассчитанный на длительные наблюдения (с 2008 по 2012 г.) и объединивший усилия лингвистов шести немецких университетов Северной Германии (Франкфурта-на-Одере, Килья, Гамбурга, Мюнстера, Билефельда и Потсдама). В связи с фактом появления разговорных форм, рассматриваемых учеными как следствие языкового контакта (взаимодействия литературного языка и нижненемецких диалектов), возникла необходимость нового исследования языковой ситуации в этом регионе: идентифицировать и описать региональные и локальные особенности языка в аспекте их территориальной, социальной и ситуативной обусловленности. Документация объективного языкового материала и фактов субъективного

восприятия (мнений, оценок, позиций, рефлексий на языковые формы) позволит создать первый представительный корпус нижненемецкого субстандарта. Исследованием охвачено 18 областей, в каждой из которых выбраны по два небольших населенных пункта (от двух до восьми тысяч жителей) и определена группа информантов (по 4 в возрасте от 40 до 55 лет), родившихся и выросших в этой местности. Анализ языковой ситуации в крупных городах, где язык подвергается воздействию со стороны многих культур, проводится отдельно в рамках проекта „Hamburgisch: Sprachkontakt und Sprachvariation im städtischen Raum“, прежде всего, в аспекте возрастающего влияния урбанизации на речевое поведение.

Любое социолингвистическое описание требует, как известно, привлечения как собственно языковых, так и метаязыковых данных, которые рассматриваются во взаимосвязи: информация о выборе языка, переключении кода (code-switching), интерференции на разных языковых уровнях и т.п., соотносится с биографиями информантов (их происхождением, условиями социализации, субъективной оценкой языковых форм, возрастными и проч. характеристиками). В названных выше актуальных проектах необходимый объем сведений собирается в различных ситуациях (Settings): интервью (для получения информации об условиях овладения языком, сферах его употребления, языковой компетенции говорящего, его оценке той или иной формы); перевод ряда стандартизированных предложений на местный диалект (для определения различий между локальными говорами); связные рассказы и описания картинок на диалекте, пробы чтения текстов на языковом стандарте, записи непринужденных бесед в семейном кругу и сбор сведений о коммуникации на рабочем месте для выявления степени использования нижненемецкой и литературной форм реализации языка в Северной Германии [6]. Так, на основании опроса в городке Бальве у южной границы региона эксперты делают вывод о пассивном владении нижненемецким в этом населенном пункте, где понимают «платтдойч», но не говорят на нём [7].

Все данные сводятся в единый электронный корпус и получают новый формат: для их обработки задействуется инструментарий, который наряду с транскрипцией собранного аудиоматериала и разметкой транскриптов и аудиосигналов дает возможность выделить релевантную для анализа информацию (аннотация). Дополнительно эта информация соотносится с обширными метаданными об информантах (биографиями) и ситуативными контекстами. Отметим, что разметка – обязательный элемент корпуса, она отличает его от простых коллекций текстов (напр., электронной библиотеки), которые можно найти в Интернете.

На основе такого корпуса возможны самые разнообразные исследования социо- и идиолектов на фонетическом, лексическом и грамматическом уровнях. Так, в центре внимания Ингрид Шрёдер (Гамбургский университет) – индивидуально обусловленные синтаксические различия [8]. Возникшие в результате языкового контакта грамматические феномены играют роль дифференциальных признаков, как в стандартной, так и в нижненемецкой формах реализации языка. Случаи нижненемецкой интерференции регистрировались в ситуациях (т.н. «ситуалектах» / *Situalekte*) с использованием языкового стандарта (чтение текста, интервью), а интерференции литературного языка – в ситуациях с использованием нижненемецкого (перевод предложений, рассказ, семейный разговор на диалекте). Т.о. в различающихся по степени официальности обстановках общения было выявлено, какие именно синтаксические «контакт-формы» реализуются в индивидуальной речи. Кроме того, были определены виды и степень языковой эрозии (*Spracherosion, language attrition*): в спектре языка, близком к стандарту, образовался (при общей тенденции к стандартизации) комплекс пространственно дифференцированных региональных языковых маркеров; в нижненемецком спектре влияние языкового стандарта обнаруживается, прежде всего, в выборе функциональных слов (устоявшихся заимствований из литературного языка).

Корпусный подход может быть положен в основу исследований языка и в какой-либо специальной области или отдельной коммуникативной ситуации для построения т.н. специальных корпусов, например, официальной речи или общения в чате (*Corpus of American and British Office Talk* или *Das Dortmunder Chat-Korpus*), которые сами становятся базой для дальнейших наблюдений. К примеру, Дортмудский корпус, созданный в период с 2002 по 2008гг. под руководством М. Байсвенгера, в открытом доступе с 2005 г. (<http://www.chatkorpus.tu-dortmund.de>). Этот корпус документирует использование языка в немецкоязычной чат-коммуникации и содержит 478 образцов из различных сфер деятельности (*Freizeit, Bildung, Beratung, Medien*) по 6000 сообщений пользователей (всего 1,06 миллионов слов текста / *Token*). При всей разработанности темы чат-коммуникации, имеются еще, как отмечает Кристин Лукхард [9, с. 1], нераскрытые аспекты, один из которых стилистическая вариативность языка. Описав многочисленные стилистические приемы экономии речевых усилий, интенсификации эмоций и солидаризации, она делает вывод о том, что специфике интернет-общения необходимо изучать не только в различных его формах (напр., в чате или в блоге), но и внутри этих форм в различных контекстах. Наряду с уже традиционными элементами (эмотиконы „:-)“ и интерактивные единицы в виде акронимов, напр., „lol“ из начальных букв „laughing out loud“; глаголы без окончаний / «инфлексивы», напр., „lach“, „seufz“), в чате используются свои характерные элементы. Со сменой его контекстов меняется и языковой стиль. Так, обращаясь к сотруднику университетской библиотеки, все пользователи отказываются от употребления эмотиконов и усеченных глаголов и ориентируются на орфографические нормы: такая интеракция закреплена за данным контекстом с его определенным набором ролей и ролевым поведением (зд., сотрудник библиотеки – читатель библиотеки). Общение между двумя лицами (диалог) продуцируется без спешки и не прерывается сообщениями других участников чата – по маркерам затраченного времени видно, что оно занимает более 5 минут (сообщения в контексте неформальной беседы пишутся гораздо быстрее) [10, S.6-8]. Корпусный подход обеспечивает, кроме того, получение количественных данных о взаимосвязи языковых (и дискурсивных) структур с медийными /техническими, контекстуальными и социальными параметрами чат-

коммуникации, которые и определяют в конечном итоге ее стиль.

Чат, сочетая в себе элементы двух способов реализации языка – письменного и устного, имеет все предпосылки для того, чтобы стать интересным объектом исследования стилистики (как «хамелеонов стиль»/ Chamäleon Stil) [9]. М. Байсвенгер отмечает, что реализация языка в интернет-коммуникации – на форумах, в чатах и блогах, по скайпу и в соц. сетях, на учебных сайтах, в компьютерных играх или «виртуальных мирах» – до сих пор еще почти не отражена в корпусах современного немецкого языка, и лишь немногие из имеющихся электронных собраний систематически документированы [11, с. 1].

Добавим к вышесказанному, что ресурсы речевых корпусов дают представительную базу и для социофонетических исследований, поскольку включают транскрипции устной речи с фонетической и просодической аннотацией и метаданными, для проведения комплексных многоуровневых исследований в фонологии; мультимедийные корпуса, содержащие текст, изображение, аудио - и видеоматериалы, к примеру, Das Archiv für Gesprochenes Deutsch (<http://agd.ids-mannheim.de>), делают возможным изучение вариативности языка в рамках устной формы разговорной речи.

Подводя **итог**, отметим, что достижения и наработки корпусной лингвистики открывают новые возможности для широкого спектра исследований теоретического и прикладного характера. Метод анализа на базе корпусов является надежным, поскольку основывается на эмпирическом подходе, использует репрезентативный объем аутентичных текстов и современные компьютерные технологии для аналитической и статистической обработки данных. Создание больших сбалансированных, лингвистически размеченных корпусов, регистрирующих различные формы реализации языка, остается сегодня актуальным. Перспективы исследований видятся как в дальнейшей разработке спектра научных тем, предполагающих использование корпусов, так и в детализации корпусно-ориентированных методов обучения иностранному языку и переводу.

Источники и литература

1. Kennedy, G. An Introduction to Corpus Linguistics / Graeme Kennedy. – London/New York: Longman, 1998. – P. 13-18.
2. Meyer, Ch. “Pre-electronic Corpora”. Corpus Linguistics. An International Handbook / Charles Meyer. – Berlin: Mouton de Gruyter, I, 2008. – P. 1-14.
3. Tognini Bonelli, E. “Theoretical Overview of the Evolution of Corpus Linguistics.” Routledge Handbook of Corpus Linguistics / Elena Tognini Bonelli. – London / New York: Routledge, 2010. – P.14-27.
4. Sinclair, John. Corpus, Concordance, Collocation / John Sinclair. – Oxford: Oxford University Press, 1991. – 179 p.
5. McEnery, R. X. Corpus-based Language Studies: an Advanced Resource Book /Richard Xiao McEnery, Tono Yukio. –New York: Routledge, 2006. – P. 3-12.
6. Lehmborg, T., Schröder I. Methoden korpusbasierter Analyse des sprachlichen Wandels in Norddeutschland / Timm Lehmborg, Ingrid Schröder [Электронный ресурс]. – Режим доступа: www.ds.uzh.ch/docs/147.pdf
7. Niemand, A. Balver Landfrauen nehmen am Projekt der Uni Bielefeld teil / Annabell Niemand // Die Meinerzhagener Zeitung, 09. November 2011 [Электронный ресурс]. – Режим доступа: <http://www.come-on.de/lokales/balve>
8. Schröder, I. „Sprachvariation in Norddeutschland“ – individuelle Aspekte der syntaktischen Variation / Ingrid Schröder [Электронный ресурс]. – Режим доступа: <http://dgfs.uni-frankfurt.de>
9. Luckhardt, Kristin. Stilanalysen zur Chat-Kommunikation. Eine korpusgestützte Untersuchung am Beispiel eines medialen Chats (2009) / Kristin Luckhardt [Электронный ресурс]. – Режим доступа: <http://hdl.handle.net/2003/26055>
10. Storrer, A. Sprachstil und Sprachvariation in sozialen Netzwerken (2011) / Storrer Angelika [Эл.рес.]. – Режим доступа: <http://www.studiger.tu-dortmund.de>
11. Beißwenger, M. Das Dortmunder Chat-Korpus: ein annotiertes Korpus zur Sprachverwendung und sprachlichen Variation in der deutschsprachigen Chat-Kommunikation (2013) / Michael Beißwenger [Электронный ресурс]. – Режим доступа: <http://www.linse.uni-due.de>