ТЕОРЕТИЧЕСКАЯ, ПРИКЛАДНАЯ И СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНАЯ ЛИНГВИСТИКА / THEORETICAL, APPLIED AND COMPARATIVE LINGUISTICS

DOI: https://doi.org/10.60797/IRJ.2024.150.41

ЛИНГВИСТИКА ДИСКУРСА И МЕТОДЫ ДИСКУРСИВНОГО АНАЛИЗА

Научная статья

Нагога О.В.[1, *]
[1] ORCID : 0000-0002-0182-3944;
[1] Саратовская государственная юридическая академия, Саратов, Российская Федерация

* Корреспондирующий автор (ok-fly[at]mail.ru)

Аннотация

Изучение дискурса является методологически сложной задачей, поскольку в настоящее время ощущается недостаток обоснованных теоретических и аналитических работ, связанных с анализом дискурса и репрезентацией его результатов. В представленной статье освещается круг вопросов, связанных с лингвистическим анализом дискурса в структурном и семантическом аспектах, дается системное и комплексное представление о научных достижениях в этой области. Рассматриваются основные вопросы определения содержания дискурсивного анализа, освещаются базовые подходы к осуществлению дискурс-анализа и приводятся различные методы цифрового анализа дискурса. Исследование создает в совокупности широкую картину, отражающую многообразие результатов современных подходов к изучению дискурса.

Ключевые слова: лингвистика дискурса, дискурсивный анализ, методы исследования дискурса, цифровой анализ дискурса, анализ ключевых слов, анализ частотных изменений, коллокационный анализ, анализ сочетаемости.

DISCOURSE LINGUISTICS AND METHODS OF DISCOURSE ANALYSIS

Research article

Nagoga O.V.[1, *]
[1] ORCID : 0000-0002-0182-3944;
[1] Saratov State Law Academy, Saratov, Russian Federation

* Corresponding author (ok-fly[at]mail.ru)

Abstract

The study of discourse is a methodologically challenging task, as there is currently a lack of grounded theoretical and analytical works related to discourse analysis and representation of its results. The presented article highlights the range of issues related to the linguistic analysis of discourse in structural and semantic aspects, provides a systematic and comprehensive view of scientific achievements in this field. The main issues of defining the content of discourse analysis are addressed, basic approaches to the implementation of discourse analysis are highlighted, and various methods of digital discourse analysis are presented. Overall, the study creates a broad picture that reflects the diversity of results of contemporary approaches to discourse studies.

Keywords: discourse linguistics, discourse analysis, discourse research methods, numerical discourse analysis, keyword analysis, frequency change analysis, collocation analysis, conjunction analysis.

Введение

Лингвистика дискурса понимается как раздел лингвистики, который изучает язык в контексте его использования в различных дискурсах и коммуникативных ситуациях. Она исследует, как формируется совокупность всех эффективных высказываний (письменных или устных), образующих тот или иной дискурс. Для анализа дискурса в целом необходима достаточная совокупность дискурсивных фактов или фрагментов дискурсов, подвергаемых анализу, однако в процессе исследования анализируемая совокупность текстов и репрезентативность их анализа могут быть поставлены под сомнение в силу ряда причин.

Заслуга ранних дискурсивно-аналитических исследований заключается прежде всего в теоретическом осмыслении и описании основной аналитической идеи исследования дискурса. Позднее, с развитием мощных вычислительных техник и носителей цифровой информации, с одной стороны, и возможностей цифровизации и, соответственно, использования цифровых технологий, с другой стороны, становится возможным добиться объективности в исследованиях дискурса.

Многообразие оцифрованных текстов дает возможность анализа репрезентативных подмножеств дискурсов. Разработанная на рубеже тысячелетий, постепенно формируется действительно новая дисциплина – цифровой дискурсивный анализ, иногда называемый дискурсивной лингвистикой, который нередко оперирует текстовыми корпусами, включающими несколько сотен миллиардов словоформ.

Методы и принципы исследования

К изучению корпусов текстов такого объема необходим особый подход, не предполагающий линейного чтения текста. Кроме того, данный подход должен иметь целью выяснить, как повторяющиеся элементы высказываний появляются в определенном количестве, распадаются, собираются вместе, расширяются, включаются в новые логические структуры или принимают новое семантическое содержание, образуя между собой языковые единства [14,

С. 89]. При этом в центре внимания проводимого анализа находится проведение количественной оценки проявления изучаемых элементов.

Шаблоны использования языка – это повторяющиеся лингвистические явления или выражения, служащие основой для определения вида и типа дискурса, которые могут быть выбраны в соответствии с различными критериями. Они могут иметь не только формальное (например, конструкции модальных глаголов), но и семантическое / лексическое сходство или составлять их комбинацию. Такой дискурс определяется на основе определенных наборов сходных моделей использования языка [10, С. 37].

Шаблон использования языка в значительной степени проявляется на лингвистическом уровне в трех наблюдаемых явлениях:

1) словосочетания, то есть случайное совместное появление двух или более выражений;

2) синтагматические шаблоны, т.е. часто повторяющиеся структуры построения аналоговых выражений;

3) частотное употребление определенных отдельных слов в текстах [6, С. 210].

Корпуса текстов сопоставляются друг с другом. Во-первых, это позволяет выявить динамику дискурса и возможные взаимозависимости элементов дискурса, а во вторых, это может быть полезным для разработки специфики конкретного тематического корпуса текстов.

Все виды анализа речевых моделей используют количественные характеристики. В процессе их применения неизбежно происходит отделение изучаемых языковых элементов от их линейного контекста. Последовательность языковых знаков нарушается, и появляются новые сочетания или контексты: поскольку лингвистические единицы вытесняются из своей линейности, их реконтекстуализация дает новый взгляд на данные [5, С. 119]. Таким образом, анализ дискурса непосредственно связан с переработкой информации, с переводом линейных лингвистических данных в табличное или иное визуальное представление данных.

Основные результаты

С момента своего появления понятие шаблона использования языка стало популярным в отечественных и зарубежных исследованиях, были созданы рабочие определения этого понятия. Однако возникает вопрос, в какой степени данный термин фактически охватывает все необходимые явления, которые могут быть эффективными критериями для дискурсивного анализа.

Так, особенности использования языка сводятся не только к использованию собственно «несущих значение» слов, но и к повторяющимся грамматическим (синтаксическим) структурам. Таким образом, «дискурсивное событие» – как расплывчато отмечает М. Фуко – выходит за семантические границы [14, С. 45].

Исследователи также рассматривают шаблон использования языка как термин, склонный к неточности. Например, Н. Бубенгофер изучает помимо семантических шаблонов формальные характеристики текстового материала, такие как средняя длина предложения, сложность текста, процентное содержание пассивных предложений и т.д. [9, С. 241-259].

Этот подход выходит за рамки собственно лингвистической прагматики с акцентом на языковые выражения, особенно если учесть, что объектом исследования является анализ дискурса не как лингвистической линейности наборов слов, а именно корпусов текстов. Это различие важно, потому что, как только язык превращается в текст, он становится больше, чем сумма его частей, «язык – это не просто набор слов, а инструмент с особыми свойствами, которые были сформированы в процессе его использования» [15, С. 156]. Анализ цифрового дискурса успешно реализуется с использованием корпусной лингвистики, которая является – в зависимости от точки зрения – субдисциплиной или методом лингвистики.

Списки ссылок на отрывки из текстов ранее создавались вручную и требовали выполнения значительного объема работы, которая, в свою очередь, должна была быть оправдана весомой идейной значимостью исходного текста, например, Библии [16, С. 5-22]. Корпусная лингвистика сегодня компьютеризирована. Производительность и память современных вычислительных машин позволяют использовать «количественные методы для выявления закономерностей использования языка в обширных текстовых корпусах для проверки соответствующих гипотез о них» [7, С. 359]. Таким образом, это помогает противодействовать «зависимости только от самостоятельного анализа интерпретатора, особенно в вопросах количественного проявления ожидаемых и/или неожидаемых лингвистических явлений» [13, С. 6].

Подобно анализу дискурса, корпусная лингвистика интересуется использованием языка, то есть конкретными лингвистическими высказываниями по теме, а не абстрактной языковой системой. Корпусная лингвистика благодаря своей количественной ориентации способна достичь уровня абстракции, который отличается от «внимательного чтения» нескольких источников и их анализа в рамках крупных тем, и, например, «беглого чтения» нескольких источников в рамках одной темы [12, С. 3].

Цель работы заключается в выявлении давно распространенных шаблонов использования языка в текстовых корпусах и их институционализации в результате лингвистико-социальных действий.

Качественные гипотезы могут быть проверены путем сопоставления их с эмпирически подтверждаемыми изменениями в языковых выражениях. Когда происходит изменение языка, это сопровождается соответствующими лингвистическими закономерностями. А если таковых нет, то существующие гипотезы подлежат пересмотру. Проверенные, а в некоторых случаях и фальсифицированные гипотезы могут привести, в свою очередь, к новым интерпретациям и к формированию новых гипотез, то есть корпусно-лингвистические исследования, с одной стороны, дают результаты, а с другой стороны, часто порождают новые вопросы.

Определение шаблонов использования языка осуществляется с помощью определенных методов анализа. Наиболее распространены четыре вида анализа: *анализ ключевых слов, анализ частотных изменений, коллокационный анализ и анализ сочетаемости.*

Анализ ключевых слов: анализ ключевых слов используется для определения того, какие слова являются существенными и несут основную смысловую нагрузку. Цель анализа состоит в том, чтобы определить тематическую или структурную специфику дискурса или его части.

Если вы хотите изучить корпус текстов, может быть полезным сравнение ключевых слов между субкорпусом и целым текстовым корпусом. Следует отметить, что в данном случае анализ ключевых слов не ограничивается только словами, несущими смысл, но также включает и знаки препинания, и сокращения, то есть все элементы, которые могут иметь значение для анализа, помимо собственно языковой семантики. Все выражения, которые распознаются компьютером как отдельные языковые субстанции, могут быть приняты во внимание при анализе ключевых слов. При проведении подобного анализа сразу бросается в глаза, что особое значение придается лексемам, которые на самом деле не являются словами в собственном смысле. Таким образом, анализ ключевых слов дискурса показывает, прежде всего, высокую релевантность интертекстуализации изучаемых языковых единиц.

Анализ частотных изменений: частотные изменения определяют относительную частоту проявления языкового объекта, его количество на миллион печатных знаков, изучаемое в определенный период времени создания текстов. При данном подходе визуализируются диахронические изменения частоты употребления словоформ, количество словоупотреблений, грамматических конструкций или словосочетаний. Кривые изменения частоты позволяют увидеть увеличение или уменьшение популярности определенных слов и словосочетаний, что, в свою очередь, позволяет сделать выводы об изменениях характеристик дискурса: увеличение частоты использования слова или словосочетания указывает на его растущую значимость в дискурсе, и наоборот. Частотный анализ приобретает особую значимость, когда несколько кривых сравниваются друг с другом.

Коллокационный анализ: значение слова определяется употреблением его в определенном контексте. Использование слова в контексте может быть проанализировано с помощью коллокаций, то есть анализа комбинаторно обусловленных лексических единиц, характеризующихся структурно-семантической целостностью. Данный метод измеряет значимость, с которой определенные слова сочетаются друг с другом или образуют сочетания слов в заданном контексте.

Если под коллокацией рассматривать перманентную сочетаемость и совместную встречаемость одних слов с другими, то при исследовании определенных терминов, часто встречающихся рядом с другими терминами, можно предположить, что между ними существует семантическая близость в рамках исследуемого текстового материала. Последовательность слов, которая встречается в корпусе текстов более одного раза в одной форме и построена грамматически корректно, анализируется далее с позиции частотности: чем выше ее значение, тем выше вероятность того, что совместное появление искомых слов не является случайным и, следовательно, имеет значимость для исследуемого текстового корпуса [11].

Анализ сочетаемости: анализ совместного появления слов в текстовом массиве больше всего напоминает выполнение поискового запроса, состоящего из одного или нескольких слов и их грамматических характеристик, представленных с одной стороны в так называемом представлении KWiC (сокращение от KeyWord in Context) и, с другой стороны, готовит их к дальнейшей обработке данных. Для небольших корпусов текстов или редких случаев употребления KWiC может стать отправной точкой для качественного подхода к анализу формы; в таком случае анализ сочетаемости служит индуктивным методом предварительной структуризации данных [13, С. 8]. Однако чем больше текстовый массив, тем менее реальным становится найти все ключевые слова, которые действительно необходимо учесть. В таком случае анализ сочетаемости служит для качественного анализа или уточнения количественных результатов. Кроме того, анализ сочетаемости может быть дополнительно переработан, и применен, например, в совокупности с частотным или коллокационным анализом.

Обсуждение

По мере развития знаний о естественном языке машинная обработка текстовых данных продолжает расширять методы цифрового анализа, используемые в корпусной лингвистике. В последние годы особое значение приобрели два метода, которые предполагают значительно более сложный анализ данных: моделирование темы LDA [1, С. 219], [2, С. 993-1022] и Word Embeddings [3, С. 562-589], [4]. Хотя оба метода, не используются в настоящей работе, представляется целесообразным представить их для описания перспектив данного исследования, и не в последнюю очередь для будущего анализа цифрового дискурса в целом.

В методе LDA (латентное размещение Дирихле) тематическое моделирование понимается как совокупность тем (Topics) и определение этих тем машинным способом. Для этого с помощью статистических методов и машинной обработки данных используются кластеры словосочетаний, то есть значимые группы словосочетаний с определенным словом, и каждый кластер представляет определенную тематическую область. Полученные таким образом темы впоследствии могут быть представлены, например, в виде кривых частотности употребления, которые позволяют определить популярность использования не одного слова, а скорее визуализировать целые тематические блоки.

Word Embeddings преследует аналогичную цель, но в первую очередь фокусируется на контекстах использования слов; это векторное представление синтаксической единицы (слова или словосочетания), позволяющее уловить контекст. По сути, эмбеддинг означает процесс или, чаще, результат процесса преобразования слов, предложений или целого текста в набор чисел – числовой вектор (массив чисел). Эти векторы обрабатываются машинными алгоритмами. Основное преимущество Word Embeddings – способность уловить семантическую сущность слов, они помогают понимать смысл и нюансы каждого слова. Кроме индивидуального значения, Word Embeddings также кодируют отношения между словами. Слова, которые часто появляются вместе в одном контексте, будут иметь похожие или «близкие» векторы, и получающийся в результате словосочетательный профиль всех конечных точек представлен в виде векторов в n-мерном пространстве. Чем ближе два вектора находятся друг к другу, тем больше сходство использования соответствующих слов и, следовательно, тем более вероятна их синонимичность или, по

крайней мере, функциональная эквивалентность. Полученные данные могут быть впоследствии дополнительно исследованы, например, с точки зрения их изменений в диахронии.

Заключение

Как показывают представленные возможности, методы анализа лингвистического дискурса в первую очередь основаны на количественных данных. Они могут описывать дискурс, но допускают лишь ограниченное понимание конкретных причин проявления определенных характеристик дискурса. Настоящее исследование сосредоточено в первую очередь на описании методов анализа дискурса и имеет своей перспективой стремление к максимально возможной информативности в работе с дискурсом с помощью комбинации различных методов исследования. Возможные интерпретации данных, полученных при помощи методов цифрового анализа, служат лишь для более глубокого понимания контекстуализации слов и не заменяют качественный историко-герменевтический анализ.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Abegg A. Empirisch-linguistische Analyse zum Wandel des Staatsverständnisses in der Schweiz / A. Abegg // Zeitschrift für Schweizerisches Recht. — 2017. — Vol. 136.

2. Blei D.M. Latent Dirichlet Allocation / D.M. Blei // Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 993–1022.

3. Bubenhofer N. Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings / N. Bubenhofer // Zeitschrift für Germanistische Linguistik. — Vol. 48. — Issue 3. — 2020.

4. Bubenhofer N. Word Embeddings: Funktionale Äquivalenz statt Synonymie / N. Bubenhofer. — 2019. — URL: https://www.bubenhofer.com/sprechtakel/2019/03/02/word-embeddings-funktionale-aequivalenz-statt-synonymie/ (letzter Zugriff: 07.10.2024).

5. Bubenhofer N. Social Media und der Iconic Turn: Diagrammatische Ordnungen im Web 2.0 / N. Bubenhofer // Diskurse – digital. — № 1. — 2019.

6. Bubenhofer N. Diskurslinguistik und Korpora / N. Bubenhofer; herausgegeben von I.H. Warnke // Handbuch Diskurs. — Berlin, New York : De Gruyter, 2018.

7. Bubenhofer N. Serialität der Singularität / N. Bubenhofer // Zeitschrift für Literaturwissenschaft und Linguistik. — 2018.

8. Bubenhofer N. Die Semantik von "Terrorismus": LDA Topic Modeling / N. Bubenhofer. — 2013. — URL: https://www.bubenhofer.com/sprechtakel/2013/03/06/die-semantik-von-terrorismus-lda-topic-modelling/ (letzter Zugriff: 07.10.2024).

9. Bubenhofer N. Korpuspragmatische Analysen alpinistischer Literatur / N. Bubenhofer, J. Scharloth // TRANEL: Travaux neuchâtelois de linguistique. — Neuchâtel : Université de Neuchâtel. — 2011.

10. Bubenhofer N. Sprachgebrauchsmuster / N. Bubenhofer // Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. — Berlin, New York : De Gruyter, 2009.

11. Bubenhofer N. Log-likelihood-Test / N. Bubenhofer. — URL: https://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=statistik_signifikanzLLR.html (letzter Zugriff: 07.10.2024).

12. Chomsky N. Aspects of the Theory of Syntax / N. Chomsky. — Cambridge : The M.I.T. Press, 1965. — 251 p.

13. Felder E. ‹Patientenautonomie› und ‹Lebensschutz›. Eine empirische Studie zu agonalen Zentren im Rechtsdiskurs über Sterbehilfe / E. Felder, J. Luth, F. Vogel // Zeitschrift für germanistische Linguistik. — 2016. — Vol. 44. — Issue 1. — S. 1–36.

14. Foucault M. Archäologie des Wissens / M. Foucault. — Frankfurt am Main : Suhrkamp, 2015.

15. Harris Z.S. Distributional Structure / Z.S. Harris // Word. — 1954. — Vol. 10. — P. 146–162.

16. Mair C. Erfolgsgeschichte Korpuslinguistik? Überlegungen zum Fortschritt in der Sprachwissenschaft / C. Mair; herausgegeben von M. Kupietz, T. Schmidt // Korpuslinguistik. — Berlin, Boston : De Gruyter, 2018. — S. 5–25.

Список литературы на английском языке / References in English

1. Abegg A. Empirisch-linguistische Analyse zum Wandel des Staatsverständnisses in der Schweiz [Empirical-linguistic analysis of the changing understanding of the state in Switzerland] / A. Abegg // Zeitschrift für Schweizerisches Recht [Journal of Swiss Law]. — 2017. — Vol. 136. [in German]

2. Blei D.M. Latent Dirichlet Allocation / D.M. Blei // Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 993–1022.

3. Bubenhofer N. Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings [Semantic equivalence in birth narratives: Application of word embeddings] / N. Bubenhofer // Zeitschrift für Germanistische Linguistik [Journal of German Linguistics]. — Vol. 48. — Issue 3. — 2020. [in German]

4. Bubenhofer Noah. Word Embeddings: Funktionale Äquivalenz statt Synonymie [Word embeddings: Functional equivalence instead of synonymy] / N. Bubenhofer. — 2019. — URL: https://www.bubenhofer.com/sprechtakel/2019/03/02/word-embeddings-funktionale-aequivalenz-statt-synonymie/ (accessed: 07.10.2024). [in German]

5. Bubenhofer N. Social Media und der Iconic Turn: Diagrammatische Ordnungen im Web 2.0 [Social Media and the Iconic Turn: Diagrammatic Orders in Web 2.0] / N. Bubenhofer // Diskurse – digital [Discourses – digital]. — № 1. — 2019. [in German]

6. Bubenhofer N. Diskurslinguistik und Korpora [Discourse linguistics and corpora] / N. Bubenhofer, edited by I.H. Warnke // Handbuch Diskurs [Manual discourse]. — Berlin, New York : De Gruyter, 2018. [in German]

7. Bubenhofer N. Serialität der Singularität [Seriality of singularity] / N. Bubenhofer // Zeitschrift für Literaturwissenschaft und Linguistik [Journal of Literary Studies]. — 2018. [in German]

8. Bubenhofer N. Die Semantik von "Terrorismus": LDA Topic Modeling [The semantics of "Terrorism": LDA Topic Modeling] / N. Bubenhofer. — 2013. — URL: https://www.bubenhofer.com/sprechtakel/2013/03/06/die-semantik-von-terrorismus-lda-topic-modelling/ (accessed: 07.10.2024). [in German]

9. Bubenhofer N. Korpuspragmatische Analysen alpinistischer Literatur [Corpus-pragmatic analyses of alpine literature] / N. Bubenhofer, Joachim Scharloth // TRANEL: Travaux neuchâtelois de linguistique [Works on linguistics in Neuchâtel]. — Neuchâtel : Université de Neuchâtel. — 2011. [in German]

10. Bubenhofer N. Sprachgebrauchsmuster [Patterns of language use] / N. Bubenhofer // Korpuslinguistik als Methode der Diskurs- und Kulturanalyse [Corpus Linguistics as a Method of Discourse and Cultural Analysis]. — Berlin, New York : De Gruyter, 2009. [in German]

11. Bubenhofer N. Log-likelihood-Test / N. Bubenhofer. — URL: https://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=statistik_signifikanzLLR.html (accessed: 07.10.2024). [in German]

12. Chomsky N. Aspects of the Theory of Syntax / N. Chomsky. — Cambridge : The M.I.T. Press, 1965. — 251 p.

13. Felder Ekkehard. ‹Patientenautonomie› und ‹Lebensschutz›. Eine empirische Studie zu agonalen Zentren im Rechtsdiskurs über Sterbehilfe [‘Patient autonomy’ and ‘protection of life’. An empirical study on agonal centres in the legal discourse on euthanasia] / E. Felder, J. Luth, F. Vogel // Zeitschrift für germanistische Linguistik [Journal of German Linguistics]. — 2016. — Vol. 44. — Issue 1. — P. 1–36. [in German]

14. Foucault M. Archäologie des Wissens [Archaeology of knowledge] / M. Foucault. — Frankfurt : Suhrkamp, 2015. [in German]

15. Harris Z.S. Distributional Structure / Z.S. Harris // Word. — 1954. — Vol. 10. — P. 146–162.

16. Mair C. Erfolgsgeschichte Korpuslinguistik? Überlegungen zum Fortschritt in der Sprachwissenschaft [The success story of corpus linguistics? Reflections on progress in linguistics] / C. Mair; edited by M. Kupietz, T. Schmidt // Korpuslinguistik [Corpus Linguistics]. — Berlin, Boston : De Gruyter, 2018. — P. 5–25. [in German]