УДК 81’11+004.9

ЭВОЛЮЦИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ И ЕЁ ПЕРЕСЕЧЕНИЕ С ГЕНДЕРНЫМИ ИССЛЕДОВАНИЯМИ: МЕТОДЫ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Василькина Е.А.

аспирант,

Мордовский государственный университет им. Н.П. Огарёва,

Саранск, Россия

Аннотация. В статье представлен анализ истории развития компьютерной лингвистики как отрасли прикладной лингвистики. Исследование выполнено в рамках междисциплинарного подхода, автором рассматривается возможность использования современного программного обеспечения для проведения исследований в области гендерной лингвистики, приведены названия, а также рассмотрены преимущества и недостатки таких приложений. В заключении даны перспективы дальнейшего исследования.

Ключевые слова: компьютерная лингвистика, машинный анализ текста, компьютерные инструменты в филологии, гендерная лингвистика, гендер.

THE EVOLUTION OF COMPUTATIONAL LINGUISTICS AND ITS INTERSECTION WITH GENDER STUDIES: METHODS AND SOFTWARE

Vasilkina E.A.

graduate student,

Mordovia State University,

Saransk, Russia

Abstract. The article presents an analysis of the history of the development of computational linguistics as a branch of applied linguistics. The research was carried out within the framework of an interdisciplinary approach, the author considers the

possibility of using modern software to conduct research in the field of gender linguistics, and provides the names, advantages and disadvantages of such applications. In conclusion, the prospects for further research are given.
Keywords: computational linguistics, machine text analysis, computer tools in philology, gender linguistics, gender.

В современном мире, где информация создается и распространяется с высокой скоростью, применение компьютерных технологий для анализа текстов становится неотъемлемой частью научных исследований. Цель настоящей статьи – рассмотреть историю становления компьютерной лингвистики и проанализировать возможность применения инструментов компьютерной лингвистики в гендерных исследованиях.

Актуальность исследования обусловлена стремительным ростом объема текстовой информации и всё более возрастающей потребностью в её эффективном анализе. Методы компьютерного анализа текста находят применение в различных областях, таких как маркетинг, медиа и лингвистика, где нужно быстро и качественно обрабатывать большие объемы данных. В связи с этим изучение методик, а также разработка новых подходов в данной области находятся в фокусе научных исследований и практической деятельности.

Компьютерная лингвистика или вычислительная лингвистика как междисциплинарная область исследований возникла на стыке лингвистики, математики, информатики и искусственного интеллекта. Развитие отрасли тесно связано с эволюцией вычислительной техники и теоретических подходов к обработке естественного языка. История компьютерной лингвистики может быть разделена на несколько ключевых этапов, каждый из которых характеризуется своими методологическими и технологическими достижениями.

Зарождение компьютерной лингвистики связано с первыми попытками автоматизированной обработки текста и машинного перевода. В 1949 году

Уоррен Уивер (Warren Weaver) предложил идею использования компьютеров для перевода текстов, что стало отправной точкой для исследований в этой области [9, 6]. В 1950-х годах в США и СССР начались активные работы по созданию систем машинного перевода. Одним из первых проектов стала система Джорджтаунского университета, которая в 1954 году продемонстрировала возможность автоматического перевода с русского на английский [6]. Однако ранние подходы, основанные на правилах (rule-based systems), столкнулись с ограничениями из-за сложности и многообразия естественных языков.

В 1960–1970-х годах развитие компьютерной лингвистики было связано с работами Ноама Хомского, который предложил теорию генеративной грамматики. Важно отметить, что Н. Хомский разграничивает обработку естественного языка (NLP), как область искусственного интеллекта, направленную на понимание текста и речи компьютерами, и лингвистику, где данные связаны с интуицией говорящего, а не с вычислительными процессами. В этот же период появились первые системы обработки текстов, такие как SHRDLU Терри Винограда, которая могла понимать и выполнять команды на естественном языке в ограниченной предметной области.

В 1960-е годы предпринимались усилия по созданию программ для разбора предложений, основанных на трансформационных грамматиках Хомского. Самый значительный и продолжительный проект в этом направлении был реализован компанией IBM [4]. Несмотря на неудачные попытки запрограммировать трансформационные грамматики, было раскрыто множество тонкостей человеческого языка, которые ранее не были так очевидны. Например, такие феномены, как полисемия, синтаксическая неоднозначность и сложные рекурсивные конструкции подтолкнули исследователей к разработке алгоритмов, способных справляться с подобными особенностями лексики. Усилия по созданию усовершенствованных алгоритмов способствовали переходу от формальной лингвистики к применению вероятностных моделей, таких как скрытые марковские модели, возникшие в 1980 годах.

С начала 2000-х годов в компьютерной лингвистике началось активное использование методов машинного обучения, включая нейронные сети. Важно отметить, что на данном этапе анализу подвергаются естественно созданные и опубликованные тексты, а не специально разработанные лингвистами для машинного анализа. В 2010-х годах с появлением глубокого обучения (deep learning) произошёл качественный скачок в обработке естественного языка. Модели на основе рекуррентных нейронных сетей (RNN) и долгой краткосрочной памяти (LSTM) позволили значительно улучшить результаты в задачах машинного перевода, анализа тональности и воссоздания текста [8].

Стоит отметить, что развитие компьютерной лингвистики не только способствовало прогрессу автоматизированного перевода, но и повлияло на создание интеллектуальных систем анализа данных. Развитие отрасли привело к созданию специальных программ, способных анализировать не только частотность употребления конкретной лексической единицы в тексте (квантитативная лингвистика), но и исследовать семантические процессы между лексическими единицами в контексте, проводить синтаксический и морфологический анализы текстов. Таким образом, современные алгоритмы могут распознавать сложные грамматические структуры и контексты, делая анализ текста более точным и независимым от человеческого фактора.

Одним из перспективных направлений применения инструментов компьютерной лингвистики являются исследования в области гендерной лингвистики, которая изучает, как языковые практики отражают, конструируют и воспроизводят гендерные различия, стереотипы и социальные роли. Компьютерная лингвистика позволяет автоматизировать процесс анализа больших текстовых массивов, выделяя гендерно-маркированные лексические единицы. Например, актуальным является исследование частотности употребления в тексте лексических единиц, связанных с традиционно «мужскими» или «женскими» ролями.

Сегодня существует сравнительно малое количество гендерных исследований, проведенных с использованием автоматической обработки текста. Преимущественная часть таких работ заключается в применении методов компьютерной обработки текста с целью создания корпуса данных и дальнейшего ранжирования по частотности употребления в корпусе текста или целого корпуса. Например, данный подход применен в исследовании Улановой Е.Э., посвященном анализу гендерных представлений в профессиональной среде синхронных переводчиков [3].

Таким образом, несмотря на преимущества использования машинного анализа текстов, программное обеспечение редко применяется в филологических исследованиях для интерпретации текстов. К числу проблем, способных привести к отказу от использования программных приложений, можно отнести отсутствие доступной инструкции по использованию программного обеспечения с целью проведения филологического анализа, сложный программный интерфейс, неструктурированность средств машинной обработки текста [2]. Не менее важной является проблема технологической совместимости и адекватности инструментов. Многие существующие программы разработаны без учета специфики филологических исследований, что делает их неудобными в использовании или непригодными для решения специфических задач. Так, программные средства, такие как LIWC (Linguistic Inquiry and Word Count), позволяют анализировать текст на уровень использования различных языковых категорий. LIWC анализирует текстовые выборки пословно и сравнивает каждую из них со словарем из более чем 2000 слов, разделенных на 74 лингвистические категории, выражая результат в процентах от общего количества слов в текстовой выборке. Некоторые категории определены чисто грамматически: например, в категории «местоимения» выполняется поиск лексических единиц по категориям «местоимения первого лица единственного числа», «местоимения первого лица множественного числа» и т.д.

По общему мнению, программное обеспечение для подсчета слов, такое как LIWC, является грубым способом изучения использования языка, так как не предоставляет возможности определить контекст или основное значение слов, также как и не способно выявить иронию или определить коннотацию слов. Данный факт определяет необходимость включения традиционных методов работы с текстом, чтобы исключить неверную классификацию лексических единиц. Например, в исследовании Gender Differences in Language Use: An Analysis of 14,000 Text Samples приведен пример неверной классификации лексической единицы «mad» в таких выражениях как «I'm mad about you» [7]. Программа определяет лексическую единицу как средство выражения гнева, несмотря на то, что в вышеприведенном контексте оно используется с явно положительной коннотацией. Тем не менее, стоит отметить, что авторы исследования определяют точность классификации программой лексической единицы «mad» как 90%.

В связи с тем, что не все программы способны провести контекстуальный анализ текста, необходимо привести примеры программных приложений, способных реализовать задачи филологического исследования. Так, для проведения исследований в рамках гендерной лингвистики представляется возможным использовать различные программы и инструменты. Так, одной из наиболее популярных программ является TextAnalyst 2.0, которая представляет собой инструмент, разработанный для выполнения семантического анализа текстов и нацеленный на автоматизацию процессов извлечения ключевых понятий и установления смысловых связей между ними. Более того, программа позволяет произвести кластеризацию информации, распределяя лексические единицы по тематическим классам [5].

Другой программой для машинного анализа текста является AntConc. К преимуществам данной программы отнесем бесплатную основу пользования, возможность работы с текстами на разных языках и достаточно широкий функционал, который позволяет выявить типичную сочетаемость слов и

частотные словоформы в загружаемом пользователем тексте [1]. Также, интерфейс AntConc интуитивно понятен и прост в использовании даже для начинающих исследователей, что позволяет быстро освоить основные функции программы.

Таким образом, компьютерная лингвистика предоставляет мощные инструменты для анализа языка, которые могут быть использованы в гендерных исследованиях. Автоматизация процесса анализа текстов позволяет исследователям работать с большими объемами данных, что делает выводы более репрезентативными и обоснованными. В условиях растущего интереса к вопросам гендерного равенства и репрезентации, компьютерная лингвистика становится важным инструментом для понимания того, как язык формирует и отражает социальные нормы и стереотипы.

Перспективы дальнейшего исследования заключаются в практическом использовании программ компьютерного анализа текста для определения гендерно-маркированных лексических единиц на материале англоязычной прозы.

Библиографический список:

1. Котюрова И.А. Корпусные исследования с помощью сервиса Antconc в условиях работы в вузе / И.А. Котюрова // Язык и культура. – 2020. – №52 [Электронный ресурс]. – Режим доступа – URL: https://cyberleninka.ru/article/n/korpusnye-issledovaniya-s-pomoschyu-servisa-antconc-v-usloviyah-raboty-v-vuze (дата обращения: 21.01.2025).

2. Сибирцева В.Г. Исследование потенциала компьютерных программ для стилистического и переводческого анализа текста и его практическое применение / В.Г. Сибирцева, Н.Х. Фролова // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2019. – №1 [Электронный ресурс]. – Режим доступа – URL: https://cyberleninka.ru/article/n/issledovanie-potentsiala-kompyuternyh-programm-dlya-stilisticheskogo-i-perevodcheskogo-analiza-teksta-i-ego-prakticheskoe-primenenie (дата обращения: 21.01.2025).

3. Уланова Е.Э. Исследование репрезентации гендера языковой личности синхронного переводчика / Е.Э. Уланова // Филология и человек. – 2024. – №3 [Электронный ресурс]. – Режим доступа – URL: https://cyberleninka.ru/article/n/issledovanie-reprezentatsii-gendera-yazykovoy-lichnosti-sinhronnogo-perevodchika (дата обращения: 21.01.2025).

4. Утробина А.А. Компьютерная лингвистика и машинный перевод: об истории становления / А.А. Утробина // Вестник Башкирск. ун-та. – 2022. – №2 [Электронный ресурс]. – Режим доступа – URL: https://cyberleninka.ru/article/n/kompyuternaya-lingvistika-i-mashinnyy-perevod-ob-istorii-stanovleniya (дата обращения: 21.01.2025).

5. Якубовский К.И. Обзор современных лингвистических технологий и систем / К.И. Якубовский, К.А. Якубовская // Вестник МГУП. – 2015. – №2 [Электронный ресурс]. – Режим доступа – URL: https://cyberleninka.ru/article/n/obzor-sovremennyh-lingvisticheskih-tehnologiy-i-sistem (дата обращения: 21.01.2025).

6. Hutchins W. J. Machine Translation: Past, Present, Future / W.J. Hutchins // Chichester [West Sussex]: Ellis Horwood, 1986. – 392 p.

7. Newman M. L. Gender Differences in Language Use: An Analysis of 14,000 Text Samples / M.L. Newman, C.J. Groom, L.D. Handelman, J.W. Pennebaker // Discourse Processes. – 2008. – №45(3) [Электронный ресурс]. – Режим доступа – URL: https://www.researchgate.net/publication/253291274_Gender_Differences_in_Language_Use_An_Analysis_of_14000_Text_Samples (дата обращения: 21.01.2025).

8. Sutskever I. Sequence to Sequence Learning with Neural Networks / I. Sutskever, O. Vinyals, Q. V. Le // Advances in Neural Information Processing Systems. – 2014. – №4 [Электронный ресурс]. – Режим доступа – URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf (дата обращения: 21.01.2025).

9. Weaver W. Translation. Memorandum / W. Weaver // Rockefeller Foundation. – 1949. – 12p. [Электронный ресурс]. – Режим доступа – URL: https://aclanthology.org/1952.earlymt-1.1.pdf (дата обращения: 21.01.2025).

Оригинальность 81%