Исследование потенциала компьютерных программ для стилистического и переводческого анализа текста и его практическое применение

В. Г. Сибирцева, Н. Х. Фролова

*Национальный исследовательский университет «Высшая школа экономики»
Нижний Новгород, Россия*

Аннотация

Предлагается критический обзор существующих компьютерных ресурсов с точки зрения их использования в филологическом анализе. В гуманитарных науках, таких как социология, психология и менеджмент, существуют различные типы программного обеспечения для многомерного анализа текста (проверка авторства, определение совместимости слов, выявление эмоциональной окраски текста и т. д.), которые могут применяться и в филологических исследованиях для интерпретации текста. Однако для филологии они остаются *terra incognita*: традиционно изучение и анализ текста в литературоведении связаны с его индивидуальной интерпретацией исследователем, поэтому до сих пор преобладает консервативный подход, не подкрепленный данными контент-анализа или статистики.

Различные программы были успешно протестированы на материале русской и иностранной художественной литературы, а в дальнейшем интегрированы в обучение студентов бакалавриата и магистратуры факультета гуманитарных наук (Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород). Накопленный пятилетний опыт позволил получить неединичные достоверные данные и выполнить сравнительный анализ компьютерных программ, таких как AntConc concordancer, многофункциональный контент-анализатор LEKTA, LF Alighner, TextAnalyst SDK. На фактическом материале доказывалась необходимость перехода от традиционного сбора художественного материала и анализа различных дискурсов к широкому применению веб-инструментов. Практическое применение компьютерных программ заполняет разрыв между современными требованиями цифровых гуманитарных наук (Digital Humanities) и консервативным подходом многих исследователей.

Авторами статьи приводятся краткие аннотации курсовых и дипломных работ, посвященных различным видам анализа текста с опорой на программные инструменты. Отдельные аспекты, посвященные перспективным, но малоиспользуемым компьютерным ресурсам, описываются подробно. При этом подчеркивается необходимость избирательного подхода к общедоступным лингвистическим корпусам, узкоспециализированным программам-конкордансам и частотным анализаторам. Применение компьютерных инструментов в анализе текста обязательно требует дальнейшей традиционной интерпретации материала. Достижение необходимого баланса в использовании программных ресурсов может означать оптимизацию анализа художественного оригинального или переводного текста наряду с определением новых критериев оценки репрезентативности и валидности результатов подобных литературоведческих и лингвистических исследований.

Ключевые слова

многомерный анализ текста, компаративный анализ, компьютерные инструменты в филологии, перевод, Digital Humanities

Для цитирования

Сибирцева В. Г., Фролова Н. Х. Исследование потенциала компьютерных программ для стилистического и переводческого анализа текста и его практическое применение // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Т. 17, № 1. С. 90–101. DOI 10.25205/1818-7935-2019-17-1-90-101

Testing Computer Programs' Potentialities in Stylistic and Translational Comparative Text Analysis

Vera G. Sibirtseva, Natalia H. Frolova

*National Research University Higher School of Economics
Nizhny Novgorod, Russian Federation*

Abstract

Traditionally, text and literature abstracts analysis is associated with the individual perception and creative interpretation of the researcher. Though many software applications are being created for multidimensional text analysis from the point of view of verification of authorship, identifying the compatibility of words and emotional coloring of the text, etc., the conservative approach prevails. No matter how many different types of software for the Humanities (such as sociology, psychology and management) there exist, they are still rarely used in philological research while interpreting a text.

The current article provides a critical review of technology enhanced philological analysis. Various computer programs were successfully tested on the material of Russian and foreign literature, and later on used to train future specialists from the Humanities department (National Research University "Higher School of Economics", Nizhny Novgorod). The accumulated experience allowed us to obtain a reliable data and perform a comparative analysis of computer programs, such as AntConc concordancer, LEKTA, LF Aligner, TextAnalyst SDK, a multifunctional content analyzer.

The article describes a survey conducted at NRU HSE to address the need to move from traditional collection of authentic materials and analysis of various discourses to the widespread use of web-based tools designed to accelerate this process and follow the contemporary requirements of the Digital Humanities. Brief prefaces of the term papers and the theses devoted to various methods of text analysis based on software tools are considered and described in detail. Specific examples emphasize the need for a selective approach to publicly available linguistic corpora and highly specialized concordance programs, as well as to tools for aligning translated texts and frequency analyzers. However, the use of computer tools in text analysis should be combined and completed with traditional methods of linguistic and stylistic interpretation. The purpose is to achieve the necessary balance in the use of software resources which can mean optimizing the comparative analysis of original or translated texts, along with the definition of new criteria for assessing the representativeness and validity of the results of literary or linguistic studies.

Though the proposed gap filling between the traditional and innovative techniques of text interpretation is rather novice, computer assisted text analysis is promising and tends to be in the mainstream of comparative linguistic and stylistic research.

Keywords

multimodal text analysis, computer tools in philological education, comparative analysis, text translation, Digital Humanities

For citation

Sibirtseva, Vera G., Frolova, Natalia H. Testing Computer Programs' Potentialities in Stylistic and Translational Comparative Text Analysis. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2019, vol. 17, no. 1, p. 90–101. (in Russ.) DOI 10.25205/1818-7935-2019-17-1-90-101

Введение

Традиционно изучение и анализ художественных произведений связаны с индивидуальным восприятием и творческой интерпретацией исследователя. В настоящее время создается много программных приложений для многомерного анализа текста. Предлагается проверка авторства, выявление сочетаемости слов и эмоциональной окрашенности текста и т. д. Тем не менее, консервативный подход в исследованиях по литературе сохраняется до настоящего времени. Для гуманитарных наук (таких как социология, психология и менеджмент) существуют различные виды программного обеспечения, однако они достаточно редко применяются в филологических исследованиях при интерпретации текста.

Идея поверить «алгеброй гармонию» не должна восприниматься буквально, компьютер не может заменить профессионального филолога. Но в процессе текстового анализа всегда появляются задачи, связанные с рассмотрением большого объема материала или с рутинным поиском однородной информации. И тогда компьютерные инструменты могут облегчить обработку материала и сэкономить время. При таком подходе к области лингвистического

и филологического (в частности, компаративного) анализа открываются новые возможности и перспективы.

Основные причины пренебрежения существующими компьютерными программами в области филологических исследований следующие: отсутствует удобный пользовательский интерфейс или пошаговые инструкции для лингвистов и филологов; средства автоматической обработки текста не структурированы с точки зрения филолога; число примеров применения программного обеспечения при анализе литературных отрывков очень малочисленно, и поэтому алгоритмы, учитывающие частотность и валидность автоматической обработки, отсутствуют.

В настоящей статье описан пятилетний опыт апробации веб-инструментов в качестве средства стилистического и лингвистического анализа целого ряда коллекций и банков различных по жанру и регистру текстов силами студентов факультета гуманитарных наук (Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород). Исследовательские проекты представляют собой курсовые и выпускные квалификационные работы, посвященные анализу содержания текстового материала с применением компьютерных программ, анализаторов и веб-инструментов. Все они были загружены в систему LMS вуза, и, следовательно, большинство из них находится в режиме ограниченного доступа. Наиболее интересные исследовательские работы студентов бакалавриата и магистратуры представлены в табл. 1.

Программные ресурсы, использованные для анализа и интерпретации художественных текстов, представляют собой две большие группы: программное обеспечение, встроенное в лингвистические исследовательские корпуса, и дополнительные компьютерные программы. Состоявшиеся защиты курсовых и выпускных квалификационных работ свидетельствуют о том, что данные компьютерные ресурсы успешно применялись студентами для определения стилистической окраски текста или его перевода, для выявления ключевых мотивов произведения, а также в исследовании компьютерных средств для изучения языка на аутентичных текстах. Таким образом, описанные в настоящей статье способы компьютерного текстового анализа могут использоваться как профессиональными лингвистами, так и новичками.

Цифровые коллекции текстов и лингвистические корпуса

Развитие цифровых технологий и интерес исследователей к стилистическим вариантам русского языка стал причиной появления цифровых лингвистически аннотированных корпусов. Работа с корпусами является на сегодняшний день одним из ведущих методов лингвистического исследования, при котором можно решить множество различных проблем. Крупным проектом, базирующимся на Национальном корпусе русского языка (НКРЯ)¹, является постоянно пополняемый раздел учебного портала НКРЯ «Проблемы русской стилистики» под руководством А. И. Левинзон [2007] и Ю. М. Кувшинской [2008]. Авторы проекта исследуют закономерности выбора лексических единиц и соответствующую частоту их использования в речи XX–XXI вв. Полное лингвистическое аннотирование текстов и удобные опции поиска облегчают сбор исследовательских данных.

В преподавании иностранных языков и литературы также могут использоваться национальные языковые корпуса как средства многомерного анализа стиля и языка автора. Однако частота использования лексических единиц не является релевантным показателем при исследовании стиля, поскольку характерный для лингвистических корпусов широкий охват всевозможных текстов (в том числе современных нехудожественных) препятствует глубокому изучению текстов одного автора.

¹ URL: <https://ruscorpora.ru/>.

Исследовательские работы студентов

Students research works

ФИО студента	Тема / Topic	Аннотация и ссылка
Одинцова Анна Николаевна	Контент-анализ переводов цикла сказок Д. Р. Кипплинга «Just So Stories» в программе ЛЕКТА / Content analysis of the translations of the number of fairy tales by D. R. Kipling “Just So Stories” via the program LEKTA	Описывается оригинальный метод анализа переводов сказок Д. Р. Кипплинга с применением программы Lekta (лексико-семантический текстовый анализатор). Результатом обработки данных является выделение факторов, определяющих неочевидные взаимосвязи в текстах. Метод позволяет оценить адекватность передачи авторских смыслов и может быть использован в практических нуждах переводоведения. https://www.hse.ru/edu/vkr/126185253
Самсонова Галина Игоревна	Компаративный анализ полисемии цветообозначений (на материале коллекций примеров русского, английского и французского языков) / Comparative analysis of the color terms polysemy (based on collections of Russian, English and French examples)	Рассматривается проблематика многозначности колоративной лексики, проводится компаративный анализ цветообозначений русского, английского и французского языков для установления особенностей, сходств и различий значений. Проводится работа с данными из корпусов текстов. Составляются триграммы, то есть наиболее вероятно встречаемые слова в окружении того или иного цветообозначения. https://www.hse.ru/edu/vkr/125370574
Внюленова Любовь Андреевна	Компьютерные средства анализа характеристик героев в произведениях В. П. Крапивина / Computer tools for analyzing the characteristics of characters from the works by V. P. Kravivin	Проводится анализ характеристик главных героев произведений В. П. Крапивина с использованием лексико-семантического анализатора ЛЕКТА. Выявленные ключевые характеристики сопоставляются с данными, полученными исследователями традиционными способами. https://www.hse.ru/edu/vkr/86916268
Молокова Светлана Вячеславовна	Составление корпуса русско-английских эквивалентов интернет-коммуникации и методика работы с ним / Building a corpus of Russian-English equivalents of Internet communication and methods of working with it	Описывается методика построения корпуса русско-английских эквивалентов интернет-коммуникации, проводится исследование некоторых понятий с применением инструментов ресурса Sketch Engine: Concordance, WordSketches, Thesaurus и Sketch-Diff. Производится оценка результатов работы корпуса, предлагаются возможные пути его дальнейшего усовершенствования и способы применения в прикладных целях. https://www.hse.ru/edu/vkr/153009761
Аджигитова Юлия Руслановна	Создание приложения «Словарь ключевых слов английских романов-антиутопий» (на материале романов Дж. Оруэлла «1984», О. Хаксли «Brave New World») / Creation of the application “Dictionary of keywords of English dystopian novels” (based on the novels by G. Orwell “1984”, O. Huxley “Brave New World”)	Выявлены основные различия и сходства рассматриваемых антиутопий. Обнаружено, что с помощью ключевых слов можно дополнить наблюдения исследователей. По итогу работы было создано веб-приложение «Словарь ключевых слов британских антиутопий», ориентированное на пользователей, знающих английский язык и интересующихся антиутопиями, филологов и лингвистов. https://www.hse.ru/edu/vkr/182595148

ФИО студента	Тема / Topic	Аннотация и ссылка
Овчинникова Наталья Александровна	Особенности контекстного анализа посредством веб-технологий / Peculiarity of Context Analysis by Means of Web-tools	Выявляются и анализируются наиболее частотные программные продукты для контекстного анализа. В связи с тем, что возможности человека-переводчика ограничены, прежде всего, по времени, предлагается возможное решение данной проблемы путем создания и развития веб-инструментов для машинного перевода. https://lms.hse.ru/?ap_work&h_id=64CCF754-7750-4847-BC97-A4942AA02B69
Овчинникова Наталья Александровна	Роль коннотации лексических единиц в контекстном анализе при работе с корпусами / Corpus Use in Context Analysis of the Connotation of the Word	Проводится уточнение значения коннотации, а также связь изучения данного явления с компьютерными технологиями. Целесообразность применения веб-инструментов показана на примере использования корпусов текстов. Целью исследования является определение роли коннотации выявления ее функций в контекстном анализе лексических единиц. https://lms.hse.ru/?ap_work&h_id=24AF64B6-1AB2-48BA-92AA-462084F80EFB
Ураев Дмитрий Юрьевич	Сравнительный анализ эффективности применения веб 2.0 инструментов в процессе изучения английского языка / Web 2.0 based tools in English learning: comparative approach	Обучение иностранным языкам все теснее связано с использованием аутентичных современных текстов и во многом обусловлено развитием компьютерных технологий и непрерывным увеличением их быстродействия. Рассматриваются и сопоставляются различные инструменты веб 2.0. https://lms.hse.ru/?ap_work&h_id=12F971D0-FDEF-44DA-AC39-4EE6565B34D4
Владимирский Арсений Ильич	Мотивы жизни, смерти и бессмертия в трагедии У. Шекспира «Гамлет» и их реализация в русских переводах / Motives of life, death and immortality in the tragedy by William Shakespeare's «Hamlet» and their implementation in Russian translations)	Проводится частотно-семантическое исследование, выделяется семантическое ядро и центр слов форм «жизнь», «смерть» и «бессмертие» и в последствии создаются лексико-семантические микрополя мотивов; анализируются переводы «Гамлета» Уильяма Шекспира на русский язык с использованием конкорданса AntConc; выявляются сходства и различия на уровне мотивов между оригиналом и переводами. https://www.hse.ru/edu/vkr/220137419
Кисляницина Елена Александровна, Сажина Алёна Сергеевна	Изучение лексики английского языка с применением информационных технологий / Vocabulary learning by means of ICT	Информационно-коммуникационные технологии призваны ускорить процесс ввода в активный оборот и запоминания новых лексических единиц из актуальных текстов, а также пополнить список современных синонимов и фразеологизмов. В печати
Манаенкова Валерия Сергеевна, Абрамова Мария Александровна	Морфологический и синтаксический состав свободных ритмов XVIII века и свободного стиха 2-й половины XX века. Традиции и новаторство / Morphology and Syntax of Free Verse in German Poetry of the 18 th Century and of the Second Half of the 20th Century: Tradition and Innovation	Проводится анализ изменения синтаксиса и морфологического состава предложений в нерифмованном немецком стихе. Ведется работа над созданием представительной коллекции текстов, выявлением характеристик стихов при помощи программных средств (конкордансер, выравниватель), проводится сопоставление стихов-оригиналов и их переводов на русский язык. В печати

Для изучения наследия отдельного автора необходимы специальные лингвистически аннотированные цифровые собрания текстов. Образцовым современным проектом, учитывающим достижения Digital Humanities и не утратившим удобство в использовании, является проект «Толстой-диджитал» [Bonch-Osmolovskaya, Kolbasov, 2015]. Сохраняя все достоинства национальных корпусов (объем материала, полное морфологическое и синтаксическое описание лексики, метаданные, интеллектуальный поиск), он обогатился мультимедийной поддержкой (аудио, фото, карты и т. д.) и доступом к образовательным мероприятиям (музеи и новости в режиме реального времени). Корпус в настоящий момент состоит из многочисленных томов известного русского писателя Льва Толстого, находящихся в процессе оцифровки и обработки [Tolstaya, 2016], ставя конечной целью создание онлайн-энциклопедии, которая позволит пользователям получать доступ ко всем текстам Толстого, по ключевым фразам находить конкретные цитаты, тексты или отрывки эпистолярного наследия писателя. Проект поддерживается ведущими мировыми университетами, включая Московский государственный университет и Гарвардский университет.

Не менее известен частотный словарь-конкорданс публицистики другого русского писателя Ф. М. Достоевского [Kopotev, 2003], созданный несколько ранее. Данный корпус текстов можно назвать удачной коллекцией готовых произведений, опирающейся на тридцатитомное академическое собрание сочинений Ф. М. Достоевского. Словарь позволяет получать контекстные и частотные характеристики любого слова или словоформы, а также находить их позицию в тексте в собрании сочинений, т. е. проводить сквозной поиск по всем текстам Достоевского. Оцифрованные и аннотированные материалы словаря-конкорданса вместе со словарем литературных текстов писателя, созданным Институтом русского языка РАН, дают полное представление о частотных характеристиках основного корпуса текстов Достоевского.

Наряду с национальными корпусами и корпусами одного автора существуют коллекции, содержащие разные варианты перевода одного текста, систематизированные и аннотированные в соответствии с конкретными целями. Так, русско-французский параллельный поэтический корпус первой трети XIX в. помогает исследовать специфику переводов нескольких авторов и анализировать оригинальные и переводные идиомы.

Параллельный сборник переводов «Слова о полку Игореве» – еще один ресурс для сравнения переводов. Возможности этого корпуса одного конкретного текста довольно широки, но прежде всего пользователь может анализировать языковые конструкции в параллельных версиях и тем самым получать полную картину восприятия «Слова о полку Игореве» в современном мире [Orekhov, Rybina, 2017].

Поскольку рассмотренные ресурсы посвящены художественным текстам, было бы логично предположить, что исследователи обращаются к данным корпусам и коллекциям постоянно. В частности, открывается перспектива изучения индивидуального стиля автора: исследование рифмы, стилистических клише и т. д.

Вышеупомянутые программы имеют удобный интерфейс, что очень важно для исследователя-филолога, однако их использование в лингвистических и литературоведческих исследованиях значительно ограничено: все они являются негибкими литературными сегментами. С помощью корпуса Ф. М. Достоевского невозможно интерпретировать произведения любого другого автора, например, Н. В. Гоголя или М. А. Булгакова. Но перспектива и актуальность таких коллекций очевидна – филологи получают хороший образец оцифрованных текстов для изучения художественных произведений с полным набором инструментальных средств: системой поиска и цитирования, лингвистической разметкой, метаданными и тезаурусом.

Программные инструменты филологического анализа

Инструмент обработки естественного языка – это специальный продукт, который может применяться при анализе художественных текстов. В некоторых исследовательских публи-

кациях содержится сводное описание инструментов, позволяющих им выполнять лингвистическую обработку текста. Однако филологи в основном избегают использования компьютерных программ в своих исследованиях [Cordell, 2015], игнорируя цифровизацию гуманитарных дисциплин (Digital Humanities). Следует отметить, что в контексте настоящей статьи рассматриваются только программы, помогающие литературоведческому анализу и интерпретации текста. За рамками статьи остаются такие программные продукты, как инструменты для фонетического и фонологического анализа текста, автоматизированные создатели словарей, инструменты автоматической индексации и извлечения терминологии.

Все существующие инструменты и программы можно разделить на несколько групп:

- 1) программы для анализа морфологических и синтаксических особенностей текстов;
- 2) программы статистического анализа текста.

Эти инструменты могут быть интересны филологам, поскольку каждый из них имеет свои преимущества. В 2001 г. появился TextAnalyst², интерфейс которого далек от современных разработок, однако он по-прежнему используется в образовательных целях, так как содержит полезные компоненты реализации комплекса функций автоматического анализа для русских и английских текстов. Программа позволяет выделить основное содержание текста, сформировать гипертекстовые базы данных, выполнить поиск информации в текстах и т. д. Интересными представляются следующие описанные функции ресурса:

- контент-анализ с гиперссылками – идентификация семантической структуры текста в виде иерархии тем и подтем;
- семантический поиск связей между запрошенными словами (с учетом скрытых семантических связей между запрошенными словами и контекстом);
- автоматически создаваемая «выжимка» текста, представляющая собой семантический портрет текста с набором ключевых слов и наиболее содержательных фраз.

Работа над проектами, посвященными филологическому анализу с использованием компьютерных инструментов, будет более эффективной при создании междисциплинарных проектных групп.

Опыт совместных исследовательских проектов (Digital Humanities projects) показывает, что применение программ для анализа литературных текстов затруднено следующими факторами: отсутствие информации о работоспособности того или иного ресурса на сегодняшний день; дороговизна и отсутствие открытого доступа к продукту; отсутствие четкой понятной инструкции для непрофессионального пользователя. Тем не менее, практическое использование некоторых веб-инструментов в настоящем исследовании продемонстрировало различные возможности для анализа, а затем и интерпретации литературных текстов с помощью современных программ.

По инициативе студентов и преподавателей ВШЭ были экспериментально проверены следующие программы: конкордансер AntConc³, многофункциональный текстовый анализатор LEKTA⁴, программа LF Aligner⁵ для выравнивания текста. Например, в процессе изучения разных переводов одного и того же текста можно отслеживать стилистические изменения в дискурсе и обнаруживать особенности стратегии перевода. Поскольку художественные тексты, как правило, содержат большое количество лакун и случаев асимметричного перевода, анализ и интерпретация произведений при помощи программных средств становится увлекательной исследовательской задачей. Несмотря на широкие возможности, указанные программы достаточно редко используются литературоведами.

Бесплатный ресурс LF Aligner можно применять для изучения несоответствий в различных переводах одного текста. Наглядным примером использования программных продуктов

² URL: <https://analyst.ru/index.php?lang=eng&dir=content/products/>.

³ URL: <https://www.laurenceanthony.net/software.html/>.

⁴ URL: <http://content-analysis.ru/>.

⁵ URL: <https://sourceforge.net/projects/aligner/>.

в литературоведении на начальных этапах может являться проект по анализу нескольких переводов на русский язык романа И. Хмелевской «Все красное» [Сибирцева, 2013], осуществленный в рамках совместной работы студентов и преподавателей. Для анализа данного произведения уже существующий лингвистический польско-русский параллельный корпус не подходит: как и рассмотренные выше национальные корпуса и коллекции текстов, он не позволяет добавлять тексты из области интересов конкретных исследователей. По сравнению с лингвистическими корпусами, программа LF Aligner для выравнивания параллельных текстов может стать дополнительным инструментом при анализе перевода по нескольким причинам. Прежде всего, визуализация нескольких вариантов перевода удобна для анализа. Во-вторых, объем программы не ограничивается двумя текстами; при необходимости могут быть добавлены четыре или более текстов на разных языках. И, наконец, удобный интерфейс программы для филолога, а также простые команды не требуют специальных навыков программирования, и весь процесс обработки текстов занимает всего несколько минут.

Разумеется, возможности программы LF Aligner намного шире, чем визуализация сравниваемых фрагментов, но наглядность в процессе сравнения может значительно облегчить восприятие нескольких текстов и сэкономить время на поиск необходимых фрагментов в каждом тексте, взятом по отдельности.

Программа для контент-анализа ЛЕКТА, разработанная в Нижегородском государственном университете и имеющая ограниченные бесплатные возможности⁶, также использовалась в анализе литературных текстов с применением компьютерных средств. Поскольку компьютерные программы дополняют друг друга и фактически усиливают любые исследования, для получения достоверных данных необходимо использовать как качественные, так и количественные методы. В лингвистике и литературоведческих исследованиях допускается применение контент-анализа. Практичность этого подхода заключается в том, что контент-анализ собирает чисто лингвистическую информацию о текстовых характеристиках и на ее основе пытается идентифицировать семантические особенности текста. Суть контент-анализа с применением компьютера – породить достоверную гипотезу о содержании текста на основе внешних (количественных) характеристик текста на уровне слов и словосочетаний и, как следствие, сделать выводы об особенностях авторского слова: намерении, реляции, системе ценностей и т. д. Использование контент-анализа в интерпретации текста накладывает определенные ограничения, о которых переводчику необходимо помнить: текстовый анализ в значительной степени формализуется, производится простая количественная регистрация языковых единиц, что позволяет прогнозировать содержание текста по семантике включенных в него единиц. Практика доказывает эффективность контент-анализа в документальной прозе, фольклоре, а также в авторских сказках. Однако тексты с большим количеством метафор, языковой игрой и сложным сюжетом не могут быть достоверно интерпретированы при помощи контент-анализа.

Контент-анализ представляет собой метод обнаружения и оценки специфических характеристик текстов и других носителей информации (например, видеороликов, телевизионных и радиопрограмм, интервью, ответов на открытые вопросы и т. д.). Представленная возможность компьютерной «кросс-обработки» текстов с дальнейшей интерпретацией результатов широко используется в социологии, позволяет идентифицировать скрытое информационное содержание и его форму. Не следует забывать, что такой анализ значительно затруднен для неподготовленного исследователя-филолога. Студент или преподаватель должен научиться работать в программе ЛЕКТА самостоятельно или с помощью специалиста.

Исследования, упомянутые в данной статье, были сделаны студентами-магистрантами после прослушивания лекций по работе с программой контент-анализа. Для изучения литературного текста с помощью программы ЛЕКТА были выбраны многочисленные переводы

⁶ См.: Рюмин А. Блог о контент-анализе. Софт. URL: <http://content-analysis.ru/index.php/luchshij-soft-dlya-kontent-analiza/>.

сказки Д. Р. Кипплинга «Как леопард стал пятнистым» [Kipling, 2013]. Авторские сказки Д. Р. Кипплинга отличаются простотой стиля и языка, план выражения непосредственно соответствует плану содержания, а лингвистические характеристики текста показывают его сходство с интервью.

Программа LEKTA способствовала полуавтоматизированному поиску ключевых слов и фраз, которые образуют точки пересечения текстовых значений. Факторный анализ (сравнение совпадающих слов) помог сделать более обоснованные выводы о семантическом аспекте текста, как поверхностном, так и скрытом. В ходе исследования переводов одной главы из сказки «Как леопард стал пятнистым» выяснилось, что многие ключевые слова связаны с цветом: например, в описании особенностей сказочной топографии (High Veld) и шкуры леопарда преобладает желтый и коричневый цвет.

В ходе исследования были не только проанализированы преимущества и недостатки одиннадцати известных переводов сказки, но и произведена компиляция гипотетически наиболее полного перевода, включающего в себя различные фрагменты авторских вариантов. Контент-анализ позволил выявить скрытый смысл текста и его переводов, а также провести полный семантический анализ сказки. Таким образом, программа для количественного социологического анализа доказала свою состоятельность в компаративных литературоведческих исследованиях. Подобный анализ представляется целесообразным для циклов коротких текстов, многозначность семантики которым не свойственна. В этом случае компьютерный поиск семантических связей может принести очень интересные результаты.

Конкордансер AntConc прост в эксплуатации и требует минимальной подготовки для текстового анализа, в связи с чем широко используется переводчиками. При этом программа вполне применима и для филологических исследований. AntConc – бесплатная программа, которая помогает применять самые простые приемы лексического анализа пользовательских текстовых коллекций, например выявлять типичную сочетаемость слов и наиболее часто встречающиеся словоформы. Неоспоримым преимуществом этой программы является способность обрабатывать общемировые языки. AntConc не может объединить различные морфологические словоформы в одну лексему (как в LEKTA), но этой проблемы можно избежать, если знать правила комбинации различных лексем. Набор инструментов AntConc довольно разнообразен, но только некоторые из них можно применять в процессе работы с литературным текстом: программа позволяет выводить согласование ключевых слов в контексте (KWIC); выбирать ключевые слова в тексте; получать частотные, алфавитно-частотные и инвертированные текстовые словари; составлять списки совпадений с искомым словом.

Данная программа применялась в интерпретации повестей В. Крапивина [2001]. В частности, для анализа стиля В. Крапивина использовались четыре текста: «Журавленок и молнии», «Лето кончится не скоро», «Бабушкин внук и его братья» и «Самолет по имени Сережка». Объектом исследования был наиболее часто встречающийся авторский типаж, так называемый «Крапивинский мальчик» [Разумихин, 1982]. Молодые исследователи попытались выявить особенности языкового поведения и характера героев Крапивина на основе коллокаций слов. Прежде всего, были проанализированы слова «мальчик», «мальчишка» и имена главных героев. С помощью частотного анализа и списков коллокаций было установлено, что в текстах этого автора часто встречаются словосочетания «весело сказал мальчик» и «мальчик улыбнулся». Книги Крапивина в основном ориентированы на подростков, и неудивительно, что имя главного героя часто сочетается с такими словами, как «обрадовался», «усмехнулся», «засмеялся», «весело», «улыбнулся»; независимо от того, какие проблемы и неприятные ситуации случаются с главным героем, автор идеализирует мир, изображая мальчика как радостного и счастливого. Тексты также содержат много описаний внешнего вида персонажа: «кудрявый мальчик», «симпатичный мальчик, рыжий», «рыжеватый мальчик», «длинноволосый мальчишка», «белобрысый мальчишка», «веснушчатый мальчишка», «мальчишка в синей рубашке», «мальчишка в широкой кепке», «желтоглазый мальчишка».

В результате анализа художественного текста, выполненного программой AntConc, была установлена корреляция между основными чертами «Крапивинского мальчика», намеченными в качестве рабочей гипотезы, и их количественными показателями. Наблюдения исследователей произведений В. Крапивина, работающих в традиционном ключе, были подтверждены статистически.

Выводы

В результате экспериментального использования программных инструментов была доказана их эффективность с точки зрения выбора стратегии перевода, анализа коллокаций или контекстуального анализа. Идея «Литературный текст vs Компьютерный анализ» представляется несостоятельной, потому что это заблуждение основано на двух изначально ложных тезисах: «компьютер способен оценить каждый оттенок смысла» и «литературные исследования не требуют компьютерного анализа вообще».

В настоящем обзоре намеренно были описаны бесплатные и частично бесплатные программы. Необходимость привлечения финансирования всегда препятствует экспериментальной исследовательской работе и является негативным фактором для исследователей (особенно студентов), подталкивающим к отказу от использования веб-инструментов. Как сказано выше, обсуждаемые программы не создавались специально для литературоведческого анализа и не стремятся полностью заменить научного исследователя. Специалисты по переводу, профессионалы в области лингвистики и литературоведы должны учитывать новые возможности цифровизации филологии. Вместе с тем данные компьютерные инструменты необходимо тестировать на новом материале, чтобы выявить возможные ограничения в их использовании. Каждая программа требует определенных навыков при работе с ней. Выводы можно найти в табл. 2.

Таблица 2

Характеристики исследуемых программ в сопоставлении

Table 2

Comparative characteristics of the programs under study

Программа Program	Большой объем текстов Great text volume	Бесплатность Free access	Удобство использования User-friendliness	Многоязычность Multilingual	Мультифункциональ- ность Multifunctioning
TextAnalyst SDK	+	+	–	–	–
LF Aligner	+	+	+	+	–
LEKTA	+	–	+	–	+
AntConc	+	+	–	+	+

В ходе экспериментального использования программ на основе тестирования были определены наиболее ценные для большинства опрошенных студентов и преподавателей характеристики представленных в табл. 2 инструментов. Выделенные преимущества:

- 1) способность обрабатывать большие массивы данных;
- 2) бесплатный доступ как для отдельных лиц, так и для организаций;
- 3) понятный интерфейс, не требующий специальных навыков;

- 4) сокращение времени работы;
- 5) возможность загрузить готовый текст без необходимости вводить его вручную.

Общим преимуществом всех подобных программ является способность обрабатывать большие объемы текстов за значительно меньшее время, чем при традиционном обращении к бумажным носителям текста.

Дигитализация всех сфер повседневной и научной жизни требует от филологов освоения новых направлений (Digital Humanities) и умения применять на практике инновационные способы интерпретации текста и анализа его содержания. Постепенно эти навыки формируются и с успехом используются, несмотря на нехватку IT-специалистов для оказания технической и просветительской поддержки в университетах. Более того, само создание языкового корпуса или учебной коллекции текстов занимает много времени, требует специализированной подготовки и скоординированной работы филологов и программистов. Но как только это происходит, у исследователей появляется много новых возможностей интерпретации слова или стиля определенного автора или его переводчика. Поэтому настоятельно рекомендуется чаще обращаться к программам, упомянутым в данной статье.

Необходимо, однако, уметь выбирать компьютерные программы и целесообразный способ контент-анализа, четко соотносить требования исследования и желание исследователя. Представленные в статье исследовательские работы являются примером данного осознанного выбора.

Область исследований на стыке компьютерных и гуманитарных наук неуклонно расширяется. Но компьютерный анализ не может заменить лингвистическую интерпретацию, проведенную филологом лично. Результаты, полученные в процессе компьютерного анализа, доказывают свою достоверность и эффективность и являются многообещающими в области литературоведческого анализа, а потому заслуживают дальнейшего внимания научной общности.

Список литературы / References

- Крапивин В. П.** Бабушкин внук и его братья // Крапивин В. П. Собр. соч.: В 30 т. М.: Центрполиграф, 2001. Т. 18. 437 с.
- Krapivin, V. P.** Babushkin vnuk i ego brat'ya. In: Krapivin, V. P. Sbornik sochinenij: In 30 vols. Moscow, Tsentrpoligraf, 2001, vol. 18, 437 p. (in Russ.)
- Кувшинская Ю. М.** Проблемы русской стилистики. 2008. URL: <https://studiorum-ruscorpora.ru/stylistics/>.
- Kuvshinskaya, Yu. M.** Problemy russkoj stilistiki [Russian Stylistics Problems]. 2008. (in Russ.) URL: <https://studiorum-ruscorpora.ru/stylistics/>.
- Левинзон А. И.** Использование НКРЯ в преподавании дисциплины «Риторика» в вузе // Национальный корпус русского языка и проблемы гуманитарного образования. М.: AST, 2007. С. 160–173.
- Levinzon, A. I.** RNC use in discipline Rithorics teaching in higher institutions. In: RNC and problems of humanitarian education. Moscow, AST Publ., 2007, p. 160–173. (in Russ.)
- Разумихин А.** Правило без исключений, или Прозрачная злость и интеллигентные мальчики Владислава Крапивина // Урал. 1982. № 8. С. 149–152.
- Razumikhin, A.** The unexceptional rule or crystal anger and decent boys by Vladislav Krapivin. *Ural*, 1982, no. 8, p. 149–152. (in Russ.)
- Сибирцева В. Г.** Технология использования параллельного подкорпуса Национального корпуса русского языка и коллекций текстов в обучении иностранным языкам // Rocznik Instytutu Polsko-Rosyjskiego. 2013. № 2 (5). URL: <https://ip-r.org/wp-content/uploads/2014/01/Sibircewa-Rocznik-IPR-2-5-2013.pdf/>.
- Sibirtseva, V. G.** Technologies of using parallel subcorpora of RNC and text collections in teaching of foreign languages. *Rocznik Instytutu Polsko-Rosyjskiego*, 2013, no. 2 (5). (in Russ.) URL: <https://ip-r.org/wp-content/uploads/2014/01/Sibircewa-Rocznik-IPR-2-5-2013.pdf/>.

- Bonch-Osmolovskaya, A., Kolbasov, M.** Tolstoy digital: Mining biographical data in literary heritage editions. In: CEUR Workshop Proceedings. 2015. URL: <https://ceur-ws.org/Vol-1399/>.
- Cordell, R.** How Not to Teach Digital Humanities. 2015. URL: <http://ryancordell.org/teaching/how-not-to-teach-digital-humanities/>.
- Kipling, J. R.** How the Leopard Got His Spots. 2013. URL: <https://etc.usf.edu/lit2go/79/just-so-stories/1304/how-the-leopard-got-his-spots/>.
- Orekhov, B., Rybina, M.** Le bestiaire médiéval dans un contexte surréaliste : à propos d'une traduction de Slovo o polku Igoreve par Philippe Soupault. In : Verba sonandi : Représentation linguistique des cris d'animaux. Presses Universitaires de Provence, 2017, p. 321–331.
- Tolstaya, F.** Tolstoy goes digital. *East-West digital news*, 2016. URL: <https://www.ewdn.com/2016/09/27/tolstoy-goes-digital/>.

Материал поступил в редколлегию
Date of submission
01.10.2018

Сведения об авторах / Information about the Authors

Сибирцева Вера Григорьевна, кандидат филологических наук, доцент департамента социальных наук Национального исследовательского университета «Высшая школа экономики» (ул. Б. Печерская, 25, Нижний Новгород, 603155, Россия)

Vera G. Sibirtseva, Associate Professor: HSE Campus in Nizhny Novgorod, Faculty of Humanities (Nizhny Novgorod), School of Social Sciences, National Research University Higher School of Economics (25 B. Pecherskaya Str., Nizhny Novgorod, 603155, Russian Federation)
vsibirtseva@hse.ru

Фролова Наталья Хайдаровна, кандидат педагогических наук, факультет гуманитарных наук, доцент департамента лингвистики и иностранных языков Национального исследовательского университета «Высшая школа экономики» (ул. Б. Печерская, 25, Нижний Новгород, 603155, Россия)

Natalia H. Frolova, Associate Professor: HSE Campus in Nizhny Novgorod, Faculty of Humanities (Nizhny Novgorod), School of Applied Linguistics and Foreign Languages, National Research University Higher School of Economics (25 B. Pecherskaya Str., Nizhny Novgorod, 603155, Russian Federation)
nfrolova@hse.ru