DOI: 10.33310/2518-7813-2019-65-2-323-328
УДК 378.147

Галина ХОДЯКОВА
кандидат педагогических наук, доцент,
доцент кафедры общей и прикладной лингвистики
Николаевского национального университета имени В. А. Сухомлинского,
г. Николаев, Украина
e-mail: khodiakovagalina@gmail.com

КОМПЬЮТЕРНАЯ ОБРАБОТКА ТЕКСТОВ В КУРСЕ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ

В статье описаны возможности современных компьютерных средств для анализа текстовой ин формации и методика их использования в процессе обучения студентов в курсе квантитативной лин гвистики. Приводятся примеры организации практической работы студентов на занятиях по темам: Частотные характеристики текста. Закон Ципфа. Семантический анализ текста. Типологические индексы Гринберга. Грамматический анализ текста, построение семантических графов. Возможно уве личение списка тем для изучения методов компьютерной обработки текстов, а также углубление зна ний путем изучения алгоритмов автоматической обработки текста.

Ключевые слова: квантитативная лингвистика, компьютерная обработка текста, анализ текста, методика, организация практической работы.

В последние годы для студентов специальности «Прикладная лингвистика» во многих высших учебных заведениях вводится учебный курс «Квантитативная лингвистика». Эта наука возникла, как раздел общей лингвистики. Она изучает и отображает лингвистические явления с помощью математических методов. Количественная оценка текстов стала эффективной и реально возможной только при широком распространении персональных компьютеров, которые способны многократно ускорить поиск, передачу и обработку информации.

Попытки оценить лингвистические явления математическими методами делались в течение многих лет. И сейчас можно уже назвать результаты, которых достигли совместные усилия многих ученых: лингвистов, математиков, специалистов в области информационных технологий и др.

Приведём примеры решенных задач:

- идентификация языка текста,
- определение автора,
- автоматическая обработка текстов в текстовых редакторах: поиск и замена слов, проверка орфографии, автотекст, автоматическое исправление опечаток;
- информационный поиск в сети Интернет;
- распознавание речи,
- голосовой ввод;
- синтез речи, чтение текстов;
- перевод речи в режиме реального времени;
- программы обучения иностранным языкам,
- голосовые помощники Siri, Алекса, O key, google, Алиса

и др.

В процессе решения задач квантитативной лингвистики применяются подходы из теории распознавания образов, математической статистики и теории вероятностей; решаются задачи искусственного интеллекта на основе нейронных сетей, кластерного анализа, построения баз знаний и многие другие.

Об автоматической обработке текстов на естественном языке и проблемах компьютерной лингвистики пишут многие ученые: Большакова Е. И., Ландэ Д. В., Клышинский Э. С., Носков А. А., Андрусенко Т. Б., Волошин В. К., Городецкий Б. Ю., Пирогова Ю. К., Широков В. А. и др. [1, 4, 5, 6, 8, 10].

Широко известны учебники по квантитативной лингвистике авторов: Арапов М. В., Баранов А. Н., Марчук Ю. Н., Хроленко А. Т. и др.

В процессе исследования нас интересовало содержание учебных курсов, которые предлагали авторы учебников.

Арапов М. В. (1988 г.) включил в учебник следующее содержание: значение количественных данных для изучения языка; частота как характеристика употребительности слова в тексте; изменчивость употребительности слова в синхронии; историческая изменчивость употребительности слова; длина слова и его употребительность; полисемия слова и его употребительность; продуктивность классов слов; однородность и регулярность отношений между единицами словаря [2].

Хроленко А. Т. (2008 г.) предлагает такой курс: корпусная лингвистика, основные направления в разработке лингвокультурологической методологии, опыт разработки комплекса лингвокультурологических методик, доминантный анализ, кластерный анализ, методика сжатия конкорданса, методика аппликации словарных статей [9].

Баранов А. Н. (2003 г.) В учебнике представлены основные направления прикладной лингвистики – компьютерная лингвистика, машинный перевод, информационно-поисковые системы, лексикография, терминоведение и терминография, методика преподавания языка, теория перевода, корпусная лингвистика, политическая лингвистика, лингвистические аспекты нейролингвистического программирования, теория воздействия [3].

Содержание раздела курса «Квантитативная лингвистика» на специальности «Прикладная и математическая лингвистика» в МГУ (2002 г.): применение различных статистических методов анализа данных в лингвистике. Основные статистические критерии проверки зависимости / независимости признаков и однородности выборок, применяемые в лингвистических исследованиях. Закон Ципфа – Мандельброта и его следствия. Квантитативные методы автоматического выделения ключевых слов и терминов. Квантитативные методы, применяемые в лексикографии. Квантитативные методы, применяемые в корпусной лингвистике. Задачи атрибуции текстов и стилеметрия.

Марчук Ю. Н. (2007 г.) Учебник посвящен лингвистическим основам обработки текстов на естественном языке посредством компьютера. Рассмотрены ввод в компьютер и обработка лингвистической информации, терминология и терминография, моделирование, экспертные системы, машинный перевод. Лингвистические проблемы. Автоматическое распознавание звуков устной речи. Распознавание изолированных слов. Распознавание графем. Исправление искаженных знаков текста. Начальная характеристика естественно-языкового текста. Диагностика искажений в словах. Лингвистическая дешифровка: Лингвистическая дешифровка как прикладная дисциплина. Статистические методы. Графематический уровень. Дериватология. и многое другое. Этот курс рассчитан на несколько семестров [7].

Для приведённых выше программ характерно использование в лингвистике целого ряда математических методов и методов математической статистики. Студенты, изучающие квантитативную лингвистику по этим программам, должны иметь глубокую подготовку в соответствующих предметных областях.

Мы полагаем, что учебный курс «Квантитативная лингвистика» мог бы быть ориентирован, прежде всего, на компьютерную обработку текстов. Такие возможности постепенно появлялись с начала 2000-х годов, когда многие исследователи и научные коллективы занимались разработкой компьютерных программ по обработке текстов. Системы «Лингвоанализатор», «Атрибутор». «Авторовед» предназначались для решения задач по определению автора текста. Всем были известны системы: Promt, Плай, Рута, ВААЛ. Promt – программа-переводчик, позволяла делать англо-русский, русско-немецкий, русско-французский переводы в обоих направлениях. Она поддерживала текстовые форматы: .txt, .doc, .rtf, .wri, .htm и была встроена в Word и Excel. Система Плай – универсальный русско-украинский переводчик. Рута – система проверки правописания и грамматики. Программа ВААЛ позволяла прогнозировать эффект воздействия текстов на массовую аудиторию, проводить углубленный контент-анализ текстов и др.

Некоторое время эти системы поддерживались и обновлялись. Но в последние годы им на смену пришли on-line сервисы, которые разрабатывают крупные корпорации Microsoft, IBM, Oracle, Google, Apple, Amazon, и известные ранее программные средства обработки текстов стали менее востребованными. Программы, появившиеся в последнее время, имеют по сравнению с их предшественниками, более высокое качество, надёжность и доступность. Их можно успешно использовать в учебном процессе.

Имея новые современные инструменты для работы с текстом, исходя из целей изучения курса квантитативной лингвистики и информационного объема учебной дисциплины, мы решили при изучении данной дисциплины сделать акцент на использование компьютера и сети Internet для проведения исследований по лингвистике.

Цель написания данной статьи – описание возможностей современных компьютерных средств для анализа текстовой информации и методики их использования в процессе обучения студентов в курсе квантитативной лингвистики.

Функциональные возможности сайтов по анализу текста

Среди значительного количества on-line программ по обработке текста мы выбрали самые популярные и надёжные в плане результата. Большая часть этих программ ориентирована на работу с русскоязычным текстом, некоторые позволяют делать также обработку текста и на других языках, в частности, на украинском, английском, немецком языках.

1. TEXT.RU (http://text.ru/) – это онлайн-сервис проверки текста на уникальность, показывает процент уникальности текста, находит дубликаты и рерайт; выполняет проверку орфографии, характеристики seo-анализа текста.

2. Istio.com (http://istio.com/rus/text/analyz/) – выполняет семантический анализ текста. Показывает такие параметры, как длина текста,

водность, тошнота, наиболее частые слова в тексте и другие параметры.

3. Lenartools.ru (http://lenartools.ru/tools/lemmatop/) – находит слова (леммы, униграммы) и словосочетания (биграммы), которые часто используются в определенном списке адресов, приводя их к основной форме слова. Например, вы можете загрузить в список ТОП–100 страниц из поисковой выдачи по любому запросу, и сервис покажет слова и словосочетания, которые часто используются в текстах ваших конкурентов. Или можете загрузить список URL на отзывы о продукте, и сервис покажет слова и словосочетания, которые волнуют пользователей по данному объекту. Использование ЛеммаТОП увеличивает качество и эффективность текстов на сайтах.

4. Сайт Главред (https://glvrd.ru/) помогает очистить текст от словесного мусора, проверяет на соответствие информационному стилю. Он подходит для рекламы, новостей, статей, сайтов, инструкций, писем и коммерческих предложений, но не подходит для стихов, художественной прозы.

5. Etxt.ru (https://www.etxt.ru/antiplagiat/) – проверка текста на уникальность. Для бесплатной версии программы доступна проверка текста объёмом до 3000 слов. Сохраняет результаты проверки на сервере и предоставляет их при необходимости.

6. Яндекс.Спеллер (https://tech.yandex.ru/speller/) помогает находить и исправлять орфографические ошибки в русском, украинском или английском тексте. Языковые модели Спеллера включают сотни миллионов слов и словосочетаний. Чтобы обнаруживать ошибки и подбирать замены, Спеллер использует библиотеку машинного обучения CatBoost, может расшифровывать искажённые до неузнаваемости слова («адникасниеи» → «одноклассники») и учитывать контекст при поиске опечаток («скучать музыку» → «скачать музыку»). Кроме того, Спеллер не «придирается» к новым словам, ещё не попавшим в словари.

7. Системы «Антиплагиат» и определение автора текста (https://content-watch.ru/text/).

Эффективный бесплатный инструмент для проверки текстов на уникальность и качество. При помощи сложного алгоритма вычитания шаблона робот определяет контентную часть, делит текст на логичные фразы и определяет, кому принадлежит их авторство. Уникальной реализацией является выделение цветом и размером слов, академическая тошнота которых в тексте выше 5%. Оптимальным значением употребления в тексте повторяющихся слов (с учетом словоформ) принято считать 3.4%. Если показатель выше, есть вероятность попадания продвигаемой страницы под фильтры поисковых систем за переоптимизацию и переспам.

8. Ресурс Адвего (https://advego.com/text/seo/) – одно из самых популярных средств seo-анализа текста. Он определяет:

- плотность ключевых слов, процент ключевых фраз;
- частотность слов;
- количество стоп-слов;
- объем текста: количество символов с пробелами и без пробелов;
- количество слов: уникальных, значимых, всего;
- водность, процент воды;
- тошноту текста, классическую и академическую;
- количество грамматических ошибок.

9. Psi-technology (https://psi-technology.net/servisfonosemantika.php) выполняет фоносемантический анализ слова, оценивая его по 25 характеристикам (фоносемантические шкалы).

10. Анализ писем (http://www.analizpisem.ru/index.html) – ешё одна программа фоносемантического анализа. Анализируемый текст или письмо должны содержать минимум специальных терминов. Анализ будет неточен на специализированных текстах, он рассчитан именно на письма личного характера или фрагменты художественных произведений, где проявляется личность автора. С помощью этой программы определяется настроение автора и даётся его характеристика по 10 параметрам.

11. Автоматическая обработка текстов (aot.ru) выполняет несколько задач:

Морфологический анализ. Можно ввести русскую, английскую или немецкую словоформу и получить нормальную форму и морфологические атрибуты, либо, по желанию, всю парадигму слова.

Синтаксический анализ. Построение синтаксических деревьев в виде графов.

Перевод с русского языка на английский. Используются результаты графематического, морфологического и синтаксического анализаторов.

Лингвистический поиск по размеченному морфологическим анализатором массиву. Можно искать по части речи и по морфологическим характеристикам. Размеченный корпус состоит из 680 миллионов слов. Поиск по леммным биграммам (54 млн.).

12. Сайт Карта слов (kartaslov.ru). Создаёт карту для заданного слова, которая показывает: значение слова, ассоциации с данным словом, синонимы, делает морфологический и морфемный разбор, показывает предложения и цитаты с данным словом.

Примеры использования современных компьютерных средств на занятиях по квантитативной лингвистике

При выполнении практических работ нами была определена общая стратегия использования компьютерных средств. Различные виды анализа текста эффективно выполняются с помощью указанных выше on-line программ. Однако анализ может быть проведён самостоятельно, в редакторе Word или с помощью электронных таблиц Excel. Для получения практических умений и проверки корректности работы программ мы вначале выполняем анализ и обработку текста в Word и/или в Excel, затем делаем тот же анализ с помощью предназначенных для этого программ и сравниваем результаты, полученные после двух этапов исследования.

Тема 1. Частотные характеристики текста. Закон Ципфа

Составление частотного словаря текста является очень распространённой задачей. С помощью него можно устанавливать связь между информационной ценностью единиц языка и количественными характеристиками слов и использовать эти данные для дальнейшей обработки и при решении конкретных задач, например, задачи атрибуции текста или для SEO-анализа текста.

Предварительно в редакторе Word студенты сами создают частотный словарь для текста из 100 слов:

- выделенный текст преобразуется в таблицу с одной колонкой;
- слова сортируются по алфавиту (в порядке возрастания);
- после сортировки определяются повторяющиеся слова и их количество;
- добавляется второй столбец, где записывается частота слова, при этом строчки с повторяющимися словами удаляются;
- далее идёт сортировка по второму столбцу (по убыванию).

Частотный словарь готов. Работа занимает 15–20 минут.

С помощью одного из on-line ресурсов: TEXT.RU, Istio.com , Адвего можно составить частотный словарь и проанализировать текст по определенным параметрам за 1–2 минуты.

Сравниваем результаты работы и делаем выводы о достоверности результатов обработки текста с помощью компьютерных программ.

На основании полученных данных можно проверять закон Ципфа о распределении частоты слов в тексте: для любого слова произведение его ранга и частоты появления будет величиной постоянной:

$$r \cdot w = c,$$

где w – частота встречаемости слова в тексте; r – ранг слова в списке; c – эмпирическая постоянная величина (коэффициент Ципфа).

Тема 2. Семантический анализ текста, Seo-анализ

Статистику текста, как и другие значимые показатели семантического анализа, можно получить, используя Word.

Определение частоты слова: в частотном словаре добавляем третий столбец, затем число повторов слова делим на общее количество слов в тексте.

Семантическое ядро – это, как правило, множество ключевых слов, которые легко найти в частотном словаре.

Стоп-слова – это слова, не несущие смысловой нагрузки (предлоги, союзы, местоимения, наиболее часто употребительные в интернете существительные, глаголы и др.).

Классическая тошнота текста

$$K = \sqrt{\omega}.$$

Академическая тошнота

$$A = \frac{\omega}{W} \cdot 100\%.$$

Здесь ω – это количество вхождений ключевого слова в тексте; W – общее число слов в тексте.

Контрольная проверка результатов делается с помощью программ TEXT.RU, Istio.com, Адвего.

Тема 3. Типологические индексы Гринберга

Основные критерии типологической характеристики языка предполагают вычисление индексов синтетичности, агглютинации, деривации, показателя словообразовательной способности языка, индексов префиксальности, суффиксальности, словоизменения.

Предварительно делается морфемный анализ текста, например, из 100 слов.

Таблица создаётся в Excel, поскольку в этом приложении удобно провести все необходимые вычисления.

Слово	Морфемная структура						Кол-во морфем	Кол-во швов
	Приставка	Корень	Соединительная гласная	Суффикс	Окончание	Постфикс		

Рисунок 1 – Оформление результатов исследования слов

Морфемную структуру слова помогает проверить ресурс kartaslov.ru.

журнал	корень
ист	суффикс
ик	суффикс
а	окончание

Рисунок 2 – Разбор слова по составу в программе kartaslov.ru

Тема 4. Грамматический анализ текста, построение семантических графов.

Суть этого способа заключается в морфологическом разборе слов и синтаксическом анализе предложений. Грамматический анализ текста удобно выполнить с помощью ресурсов kartaslov.ru, aot.ru.

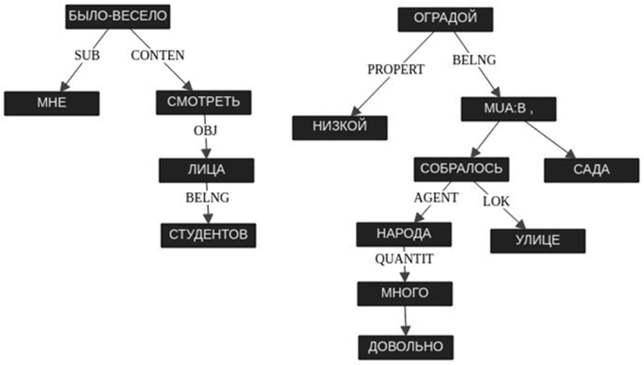

Рисунок 3 – Построение семантических графов в aot.ru

В программе aot.ru есть список семантических отношений. С помощью него можно определить отношения между вершинами графа.

Например,

SUB – субъект/подлежащее;

CONTEN – содержание;

OBJ – объект;

BELNG – принадлежность чему-то/кому-то;

PROPERT – признак чего-то/кого-то;

AGENT – агент;

LOK – локация;

QUANTIT – количество;

MUA:B – обе позиции выражают одинаковые отношения.

На практических занятиях компьютерная обработка текстов выполняется также при фоносемантическом анализе слов и текста, для идентификации автора произведения, при нахождении объема информации в лингвистической единице.

В программе обучения студентов на специальности «Прикладная лингвистика» на изучение дисциплины «Квантитативная лингвистика» отводится 3 кредита, 10 лекционных часов и 20 часов – практических занятий.

В перспективе возможно расширение учебного курса квантитативной лингвистики, увеличение списка тем для изучения методов компьютерной обработки текстов, а также углубление знаний путем изучения алгоритмов автоматической обработки текста. Это может быть предметом дальнейшего исследования в области преподавания курса квантитативной лингвистики.

Список использованной литературы

1. Андрусенко Т. Б. Лингвистические структуры в компьютерных средах. Киев, 1994. 160 с.
2. Арапов М. В. Квантитативная лингвистика. М.:Наука, 1988. 184 с.
3. Баранов А. Н. Введение в прикладную лингвистику. М., 2003. 347 с.
4. Большакова Е. И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. М.: МИЭМ, 2011. 272 с.
5. Волошин В. К. Комп'ютерна лінгвістика: навчальний посібник. Суми: Університетська книга. 2004. 381 с.
6. Городецкий Б. Ю. Компьютерная лингвистика: моделирование языкового общения // Новое в зарубежной лингвистике. М., 1989. – Вып. 24.
7. Марчук Ю. Н. Компьютерная лингвистика. М., 2007. 317 с.
8. Пирогова Ю. К. Рекламный текст: Семиотика и лингвистика. М., 2000. 263 с.
9. Хроленко А. Т. Основы лингвокультурологии: учебное пособие. М., Флинта : Наука, 2008. 184 с.
10. Широков В. А. Феноменологія лексикографічних систем. Київ:Наукова думка, 2004. 327 с.

References

1. Andrusenko T. B. (1994). *Lingvisticheskie strukturyi v kompyuternyih sredah [Linguistic structures in computer environments].* Kiev, 160 [in Russian].
2. Arapov M. V. (1988). *Kvantitativnaya lingvistika [Quantitative linguistics].* Moscow:Nauka, 184 [in Russian].
3. Baranov A. N. (2003). *Vvedenie v prikladnuyu lingvistiku [Introduction to Applied Linguistics].* Moscow, 347 [in Russian].
4. Bolshakova E. I. (2011). *Avtomaticheskaya obrabotka tekstov na estestvennom yazyike i kompyuternaya lingvistika: uchebnoe posobie [Automatic Natural Language Processing and Computational Linguistics: Tutorial].* Moscow:MIEM, 272 [in Russian].
5. Voloshin V. K. (2004). *Komp'yuterna lIngvIstika: navchalniy posIbnik [Computer linguistics: tutorial].* Sumy: UnIversitetska kniga. 381 [in Ukrainian].
6. Gorodetskiy B. Yu. (1989). *Kompyuternaya lingvistika: modelirovanie yazyikovogo obscheniya [Computational linguistics: modeling language communication]*// Novoe v zarubezhnoy lingvistike. Moscow, Vyip. 24 [in Russian].
7. Marchuk Yu. N. (2007). *Kompyuternaya lingvistika [Computer linguistics].* Moscow, 317 [in Russian].
8. Pirogova Yu. K. (2000). *Reklamnyiy tekst: Semiotika i lingvistika [Promotional Text: Semiotics and Linguistics].* Moscow, 263 [in Russian].
9. Hrolenko A. T. *Osnovyi lingvokulturologii: uchebnoe posobie [Basics of linguoculturology: study guide].* Moscow: Flinta : Nauka, 184 [in Russian].
10. Shirokov V. A. *FenomenologIya leksikografIchnih sistem [Phenomenology of lexicographic systems].* – K.: Naukova dumka, 2004. – 327 [in Russian].

Галина Ходякова. Комп'ютерна обробка текстів в курсі квантитативної лінгвістики

У статті описується один з підходів до викладання курсу квантитативної лінгвістики. Незважаючи на те, що курс порівняно молодий, вже є певні традиції в його викладанні. Зазвичай акцент робиться на використання ряду математичних методів і методів математичної статистики. Студенти, які вивчають квантитативну лінгвітстіку за цими програмами, повинні мати глибоку підготовку у відповідних предметних областях.

З початку 2000-х років ведеться активна розробка комп'ютерних програм по обробці текстів і є приклади використання цих програм в навчальному процесі. В даний час підтримка і оновлення створених раніше програм не актуальні, оскільки для аналізу і обробки лінгвістичної інформації активно використовуються on-line сервіси, які розроблені великими корпораціями. Програми, що з'явилися останнім часом, мають порівняно з їх попередниками, більш високу якість, надійність і доступність. Їх можна успішно використовувати в навчальному процесі.

Мета написання даної статті – опис можливостей сучасних комп'ютерних засобів для аналізу текстової інформації і методики їх використання в процесі навчання студентів в курсі квантитативної лінгвістики.

У статті описані функціональні можливості ряду популярних on-line сервісів з обробки та аналізу тексту. Далі наводяться приклади організації практичної роботи студентів на заняттях за темами: Частотні характеристики тексту. Закон Ціпфа. Семантичний аналіз тексту. Типологічні індекси Грінберга. Граматичний аналіз тексту, побудова семантичних графів.

Комп'ютерна обробка текстів використовується також при фоносемантичному аналізі слів і тексту, для ідентифікації автора твору, при знаходженні обсягу інформації в лінгвістичної одиниці.

У програмі навчання студентів на спеціальності «Прикладна лінгвістика» на вивчення дисципліни «Квантитативна лінгвістика» відводиться 3 кредиту, 10 лекційних годин і 20 годин практичних занять.

У перспективі можливе розширення навчального курсу квантитативної лінгвістики, збільшення списку тем для вивчення методів комп'ютерної обробки текстів, а також поглиблення знань шляхом вивчення алгоритмів автоматичної обробки тексту. Це може бути предметом подальшого дослідження в області викладання курсу квантитативної лінгвістики.

Ключові слова: квантитативна лінгвістика, комп'ютерна обробка тексту, аналіз тексту, методика, організація практичної роботи.

Galina Khodiakova. Computer processing of texts in quantitative linguistics course

In the article one of the approaches to teaching quantitative linguistics course is described. In spite of the fact that the course is relatively new there are already some traditions in its teaching. Usually the usage of a number of mathematical methods and methods of mathematical statistics is accented. Students studying quantitative linguistics according to these programs are required to have a deep understanding in the corresponding subject areas.

From the beginning of the 2000s text processing computer programs have been actively developed, there are examples of using these programs in studying process. Supporting and updating previously created programs is not of current interest, online services developed by big corporations are widely used for analysis and processing of linguistic information. Programs that appeared lately have much better quality, reliability and availability compared to their predecessors. They can be successfully used in studying process.

The goal of writing this article is to describe the possibilities of modern computer means for text information analysis and the methods of their usage in the process of teaching students the course of quantitative linguistics.

The functionalities of a number of popular online text processing and analysis services are described in this article. Further in the article examples of the practical work on following topics are given: Text frequency characteristics, Zipf's Law, Semantic text analysis, Typological indices of Greenberg, Grammar text analysis, building semantic graphs. Computer text processing is used also during phonosemantic analysis of words and text, identification of the author of a text, finding the amount of information in the linguistic unit.

In the program of teaching students on specialization «Applied linguistics» for studying the discipline «Quantitative linguistics» 3 credits, 10 hours of lectures and 20 hours of workshops are allocated.

In prospect, the development of quantitative linguistics teaching course, extension of a list of topics for studying methods of computer text processing, deepening knowledge by studying algorithms of automated text processing are possible. This can be a subject for further research in the field of teaching quantitative linguistics course.

Keywords: quantitative linguistics, computer text processing, text analysis, methods, organizing of practical work and workshops.