Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 0005.311.121:004.622

Джаримбетов

Тимурбек Байымбет улы

# Статистические методы обработки текстовых данных

**АВТОРЕФЕРАТ**
на соискание степени магистра

по специальности 7-06-0612-01 «Программная инженерия»

Научный руководитель

Хмелева А.В.

к.т.н., доцент

Минск 2025

# ВВЕДЕНИЕ

Современная эпоха цифровой трансформации характеризуется экспоненциальным ростом объемов текстовых данных, создаваемых в различных областях – от научной деятельности до повседневной коммуникации. Профессия программиста на июнь 2025 года продолжает оставаться одной из наиболее востребованных, что подтверждается прогнозами аналитических агентств, таких как Gartner, прогнозирующих рост рынка ИТ-услуг на 6,2% в текущем году. В условиях такой динамики изучение профессиональных ориентаций студентов технических вузов, включая Белорусский государственный университет информатики и радиоэлектроники (БГУИР), приобретает особую актуальность, поскольку восприятие будущей профессии напрямую влияет на мотивацию, качество подготовки и выбор карьерного пути.

Для обработки и анализа больших объемов данных необходимы методы, способные выделять значимые закономерности и устанавливать смысловые связи на основе количественных данных, что подчеркивает потребность в автоматизированных подходах. Традиционные методы анализа, такие как опросы и анкетирование, ограничены субъективностью респондентов и недостаточной глубиной лингвистического анализа, что снижает их эффективность в условиях больших текстовых массивов. В противовес этому корпусная лингвистика и методы обработки естественного языка (NLP) предлагают объективные инструменты для выявления скрытых закономерностей в текстах. Среди них особое значение имеет анализ коллокаций – устойчивых словосочетаний, отражающих концептуальные и семантические связи, которые играют ключевую роль в раскрытии тематической направленности текста. Для русского языка, отличающегося богатой морфологией и гибким порядком слов, адаптация этих методов представляет собой сложную, но перспективную задачу, подчеркивающую научную новизну данного исследования.

Настоящая работа посвящена разработке и применению методики автоматического извлечения и анализа коллокаций типа «прилагательное + существительное» в корпусе эссе студентов БГУИР, написанных на тему «Я — программист». Цель исследования заключается в выявлении ключевых аспектов восприятия профессии программиста через статистический анализ текстов. Актуальность исследования обусловлена возрастающим значением профессии программиста в условиях цифровой трансформации общества, ростом ИТ-образования в Беларуси (увеличение числа студентов на 15% в 2024-2025 учебном году), недостаточной изученностью профессионального дискурса студентов и практической значимостью результатов для оптимизации образовательных программ. Объектом исследования является корпус текстов эссе студентов БГУИР, а предметом — статистические

особенности этих текстов, связанные с восприятием профессии. Методологическую основу составляют методы корпусной лингвистики, статистический анализ и инструменты NLP, адаптированные к специфике русского языка. Работа включает теоретический обзор, описание разработанного метода с этапами подготовки данных и используемыми инструментами, что обеспечивает системный подход к решению поставленных задач.

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Диссертационная работа посвящена изучению лексических особенностей дискурса, отражающего представления студентов о профессии программиста, посредством статистического анализа коллокаций типа «прилагательное + существительное».

**Актуальность темы исследования.** Современная информационная среда характеризуется экспоненциальным ростом объемов неструктурированных текстовых данных, требующих эффективных методов анализа. В контексте формирования профессионального самосознания будущих инженеров-программистов, особую значимость приобретает изучение того, как языковые средства отражают и формируют образ профессии. Отсутствие системных исследований, посвященных комплексному корпусному анализу русскоязычных студенческих эссе на тему IT-специальностей, создает значительный пробел в понимании актуальных представлений о данной сфере. Современные методы корпусной лингвистики и автоматической обработки естественного языка (NLP) обладают значительным потенциалом для решения этой проблемы, позволяя выявлять скрытые закономерности в больших массивах текстов, минимизируя субъективность интерпретации.

**Степень разработанности проблемы.** Проблема анализа профессионального дискурса активно исследуется. Фундаментальные основы для анализа текстовых данных заложили работы по корпусной лингвистике (Д. Бибер, М. Халлидей, Д. Лич), статистической лингвистике (Г.А. Козлов, Н.А. Шехтман) и контент-анализу (Б. Берельсон, О.В. Крылова). Развитие программных средств для обработки естественного языка (таких как NLTK, PyMorphy2) значительно расширило возможности эмпирических исследований. Однако комплексный подход к изучению коллокаций типа «прилагательное + существительное» в русскоязычных студенческих эссе, посвященных профессии программиста, с целью выявления доминирующих представлений о ней, является недостаточно исследованным направлением, что создает научный пробел.

**Цель исследования.** Создание и использование метода определения и анализа коллокаций вида «прилагательное + существительное», отражающих

образ профессии инженера-программиста в сознании студентов Белорусского государственного университета информатики и радиоэлектроники (БГУИР), для углубления понимания языковых механизмов формирования профессионального самосознания.

**Задачи исследования:**

- Сформировать специализированный корпус студенческих эссе, репрезентативный для данной группы респондентов.
- Разработать алгоритм и программное обеспечение для автоматизированной предварительной обработки текстовых данных (токенизация, лемматизация, морфологическая разметка).
- Осуществить автоматизированное извлечение коллокаций типа «прилагательное + существительное» из сформированного корпуса.
- Провести количественный анализ выявленных коллокаций с применением различных статистических мер ассоциации (PMI, t-критерий, LLR) для оценки силы и значимости их связей.
- Выполнить качественный и семантический анализ наиболее значимых коллокаций для выявления их смысловых оттенков и коннотаций.
- Разработать и применить систему тематической классификации коллокаций с целью определения доминирующих аспектов образа профессии программиста в сознании студентов.
- Визуализировать полученные результаты для их наглядного представления и облегчения интерпретации.

**Объект исследования.** Лексические особенности дискурса, отражающего представления студентов о профессии программиста.

**Предмет исследования.** Коллокации типа «прилагательное + существительное» как ключевые индикаторы формирования образа профессии в студенческих эссе, а также методы их автоматизированного извлечения, статистического и семантического анализа.

**Область исследования.** Содержание диссертации соответствует образовательному стандарту высшего образования второй ступени (магистратуры) специальности 7-06-0612-01 «Программная инженерия». Ключевыми объектами исследования выступают процессы автоматизированной обработки текстовых данных, включая этапы предварительной подготовки, извлечения и анализа лингвистических единиц, а также статистические методы и алгоритмы, используемые для выявления и интерпретации закономерностей в этих данных, на примере коллокаций, отражающих профессиональные представления.

**Теоретическая и методологическая основа исследования.** В диссертации использованы исследования отечественных и зарубежных ученых, посвященные корпусной лингвистике, статистическим методам анализа текста, психолингвистике и когнитивной лингвистике. Теоретическую основу исследования составляют положения о взаимосвязи языка и мышления, концепции языковой картины мира, теории коллокации и принципы контент-анализа. Методологической основой служат достижения в области корпусной лингвистики, статистической лингвистики и автоматической обработки естественного языка. В работе применяются общенаучные методы (системный подход, сравнительный анализ, метод формализации), а также специализированные лингвистические и статистические методы (частотный анализ, меры ассоциации PMI, t-критерий, LLR, тематическое моделирование).

**Информационная база работы.** Сформирована на основе массива русскоязычных студенческих эссе (БГУИР), научной литературы по корпусной, статистической и прикладной лингвистике, а также открытых лингвистических ресурсов и библиотек для обработки естественного языка.

**Инструментальная база.** Включает программные средства и библиотеки для работы с текстом на языке Python: NLTK (токенизация, стоп-слова), PyMorphy2 (лемматизация, POS-теггинг), re (регулярные выражения), collections (Counter, defaultdict), math (логарифмы), matplotlib.pyplot (визуализация).

**Научная новизна и значимость полученных результатов.** Заключаются в разработке и применении комплексной методики корпусного анализа для определения и интерпретации коллокаций «прилагательное + существительное» в специфическом дискурсе студенческих эссе о профессии программиста. Новаторским является интеграция подробного предварительного этапа планирования эксперимента (адаптированного под корпусное исследование), многомерного статистического анализа мер ассоциации и целевой тематической классификации, что позволило получить не только количественные, но и глубокие качественные выводы о доминирующих представлениях об IT-профессии.

**Основные положения, выносимые на защиту:**

1. Разработанная комплексная методология корпусного анализа, включающая этапы предварительной обработки текста, извлечения коллокаций, многомерного статистического анализа мер ассоциации и тематической классификации, позволяет эффективно выявлять и интерпретировать специфические языковые закономерности в профессионально ориентированном дискурсе.

2. Применение статистических мер ассоциации (PMI, t-критерий, LLR) в сочетании с качественным семантическим анализом позволяет объективно оценить степень устойчивости коллокационных связей и раскрыть смысловые оттенки, формирующие образ профессии программиста в сознании студентов.

3. На основе анализа коллокаций «прилагательное + существительное» в корпусе студенческих эссе выявлены доминирующие тематические аспекты восприятия профессии программиста (финансовые, характер работы, навыки и знания, востребованность, личные качества), что предоставляет эмпирически обоснованные данные для понимания мотиваций и ценностных ориентиров будущих IT-специалистов.

4. Созданный программный инструментарий, основанный на Python-библиотеках NLTK, PyMorphy2 и Matplotlib, демонстрирует эффективность и гибкость для автоматизированного анализа текстовых данных в лингвистических исследованиях.

**Теоретическая значимость диссертации.** Заключается в уточнении методологических основ корпусной лингвистики применительно к анализу узкоспециализированных дискурсов, а также в развитии теоретических представлений о роли коллокаций как индикаторов профессионального самосознания.

**Практическая значимость работы.** Состоит в том, что полученные результаты могут быть использованы для: модернизации учебных программ и курсов по профориентации в IT-сфере; разработки рекомендаций для HR-специалистов IT-компаний по пониманию ожиданий и мотивации молодых специалистов; формирования специализированных лексических ресурсов; создания основы для дальнейших исследований в области анализа профессионального дискурса и студенческой речи.

**Апробация результатов исследования.** Результаты исследований, вошедшие в диссертацию, были опубликованы в электронном сборнике материалов 61-й научной конференции аспирантов, магистрантов и студентов БГУИР (г. Минск, Беларусь, 2025 год), в научном журнале «Студенческий вестник» (Интернаука 2025. № 17(350) часть 8).

**Структура и объем диссертации.** Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников и приложений. Общий объем работы составляет 74 страниц. Работа содержит 3 таблицы, 9 рисунков. Список литературы включает 30 наименований, список собственных публикаций соискателя из 2 наименований, 2 приложения.

# КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

**Во введении** обоснована актуальность темы диссертации, определены основные направления исследований, актуальность задачи разработки имитационных моделей для систем тропосферной связи с учетом взаимодействия «оператор - система». Также сформулированы цель и задачи диссертационных исследований.

**В первой главе «Теоретические основы коллокационного анализа»** всесторонне определено понятие коллокации как устойчивого словосочетания, элементы которого демонстрируют статистически значимую тенденцию к совместной встречаемости. Проанализированы их ключевые лингвистические характеристики, включая семантическую мотивированность, частичную предсказуемость, различия в валентности и различную степень идиоматичности. Отдельное внимание уделено разграничению лексических и грамматических коллокаций, а также подробному рассмотрению коллокаций типа «прилагательное + существительное», что соответствует фокусу последующего практического исследования.

Подробно рассмотрены основополагающие методы статистической оценки коллокаций: показатель взаимной информации (PMI), t-критерий и отношение правдоподобия (LLR). Проанализированы преимущества и ограничения каждой меры, что позволило обосновать выбор наиболее релевантных статистических инструментов для анализа корпуса студенческих эссе. Детально описаны принципы статистического подхода к выделению коллокаций, охватывающие этапы от формирования и предварительной обработки данных (токенизация, лемматизация, POS-теггинг, фильтрация стоп-слов) до определения контекстного окна и статистической оценки. Обсуждены вызовы, связанные с анализом русского языка (например, свободный порядок слов и морфологическая сложность), а также ограничения статистических методов для малых корпусов и проблема разреженности данных. Подчеркнута роль визуализации в интерпретации результатов и кратко упомянуты современные тенденции в коллокационном анализе с использованием методов машинного обучения. Таким образом, данная глава не только систематизировала актуальные теоретические данные в области корпусной лингвистики и коллокационного анализа, но и заложила надёжный методологический фундамент для разработки практической методики выявления и изучения коллокаций в специфическом языковом материале.

**Во второй главе «Методика проведения анализа и реализация алгоритмов»** детально представлена разработанная и реализованная методика анализа коллокаций. Обоснован выбор программных инструментов (Python версии 3.10.7, NLTK, PyMorphy2, collections, re, math, matplotlib.pyplot) и описаны ключевые этапы обработки корпуса эссе студентов БГУИР. Корпус

объемом приблизительно 36953 слова (110 эссе) является специализированным и тематически однородным, посвященным профессии программиста. Процесс сбора данных заключался в агрегации текстовых файлов эссе (например, из файла esse.txt) в единый корпус.

Этапы предобработки включают: очистку и нормализацию текста (удаление служебных символов, приведение к нижнему регистру, удаление избыточных пробелов и пунктуации), токенизацию (разделение на предложения и слова), морфологическую аннотацию (лемматизация с PyMorphy2 и POS-теггинг), а также удаление стоп-слов. Эти шаги обеспечили формирование высококачественного лингвистического корпуса.

Алгоритм извлечения и статистической оценки коллокаций включает: поиск целевых биграмм («прилагательное + существительное») в пределах контекстного окна до двух промежуточных слов (расстояния 0, 1, 2); подсчет абсолютных частот биграмм и отдельных слов; расчет статистических мер ассоциации – t-критерия, PMI и LLR. Для обеспечения релевантности и точности извлекаемых коллокаций применялись методы фильтрации: использование контекстных окон (с подтверждением оптимальности окна в два слова), пороговые значения для частотности (не менее 5 раз) и статистических показателей (t-тест > 1.96, PMI > 0), а также исключение шумовых данных на основе семантической релевантности к IT-тематике. Визуализация данных осуществлялась с помощью matplotlib.pyplot для представления частот, значений мер ассоциации и тематической классификации, что облегчило интерпретацию результатов.

**В третьей главе «Интерпретация результатов анализа»** представлены и проанализированы эмпирические данные, полученные в ходе исследования. Общая характеристика исследуемого корпуса (36953 слова из 110 эссе) подтверждает его тематическую однородность.

Результаты извлечения коллокаций показали значительное количество уникальных биграмм «прилагательное + существительное». В таблице ниже приведены наиболее частотные и статистически значимые из них (по общей частоте встречаемости), а также значения PMI, t-критерия и LLR:

| Биграмма (прилагательное, существительное) | Общая частота | PMI | t-критерий | LLR |
|---|---|---|---|---|
| программный, обеспечение | 76 | 9.76 | 256.04 | 1179,85 |
| первый, очередь | 46 | 9.89 | 208.91 | 670.68 |
| программный, инженерия | 40 | 9.86 | 192.85 | 573,62 |
| заработный, плата | 34 | 12.08 | 383.00 | 637.19 |
| современный, мир | 30 | 8.52 | 104,84 | 345.49 |
| точный, наука | 26 | 11.25 | 251.75 | 457.58 |

| | | | | |
|---|---|---|---|---|
| интересный, проект | 24 | 7.83 | 73,51 | 234.12 |
| различный, язык | 20 | 6.40 | 40.57 | 147.77 |
| различный, программирование | 20 | 5.39 | 28.29 | 118.63 |
| карьерный, рост | 20 | 11.15 | 213.39 | 337.03 |
| карьерный, лестница | 20 | 11.50 | 240,76 | 346.68 |
| разный, сфера | 20 | 5.94 | 34.48 | 134,95 |
| первый, программист | 18 | 4.36 | 18.27 | 79.70 |
| необходимый, знание | 18 | 7.15 | 50.24 | 156.25 |
| собственный, проект | 18 | 8.75 | 87.68 | 204.86 |
| огромный, количество | 18 | 10.26 | 148.25 | 269,93 |
| новый, язык | 18 | 5.39 | 26.79 | 104.51 |
| крупный, компания | 18 | 9.18 | 101.94 | 233,67 |
| разный, программирование | 18 | 5.38 | 26.74 | 106.34 |
| программный, продукт | 16 | 8.85 | 85,64 | 194.24 |
| первый, курс | 16 | 7.49 | 53.33 | 144.48 |
| новый, программирование | 16 | 4.21 | 16.29 | 66.11 |
| разный, задача | 16 | 6.68 | 40.05 | 123,88 |
| разный, язык | 16 | 6.22 | 34.04 | 113.03 |
| базовый, знание | 16 | 8.57 | 77.69 | 192.09 |
| высокооплачиваемый, работа | 16 | 6.61 | 39.08 | 129.33 |
| правильный, решение | 14 | 8.42 | 69.13 | 150.42 |
| новый, цель | 14 | 4.76 | 18.78 | 68.28 |
| профессиональный, задача | 14 | 7.34 | 47.35 | 123,92 |
| профессиональный, деятельность | 14 | 8.51 | 71.27 | 149,51 |

Анализ представленных данных показывает, что среди наиболее частотных и статистически значимых коллокаций выделяются сочетания, напрямую связанные с предметной областью IT (например, «программное обеспечение», «программная инженерия», «программный продукт», «новый язык», «различные языки»). Значительное количество коллокаций отражает также общие аспекты профессиональной деятельности и личные устремления, такие как «заработная плата», «карьерный рост», «карьерная лестница», «интересный проект», «высокооплачиваемая работа», «собственный проект», «необходимые знания».

Также были проанализированы частотные характеристики отдельных частей речи. Ниже представлена таблица Топ-10 высокочастотных существительных и прилагательных:

| Высокочастотные существительные | Частота | Высокочастотные прилагательные | Частота |
|---|---|---|---|
| программист | 454 | новый | 116 |
| программирование | 273 | программный | 75 |
| работа | 241 | первый | 71 |
| сфера | 206 | высокий | 65 |
| цель | 163 | различный | 64 |
| жизнь | 141 | интересный | 60 |
| язык | 136 | разный | 58 |
| знание | 129 | будущий | 38 |
| обучение | 129 | современный | 36 |
| задача | 99 | необходимый | 36 |

Среди существительных доминируют термины, напрямую связанные с IT («программист», «программирование», «язык», «знание», «задача»), а также общие понятия, важные в контексте самоопределения («работа», «сфера», «цель», «жизнь», «обучение»). Прилагательные охватывают как оценочные характеристики («новый», «высокий», «интересный»), так и описательные («программный», «различный», «необходимый»).

Помимо статистических мер ассоциации, был проведен анализ семантической близости биграмм, который дает представление о концептуальной взаимосвязи слов вне их непосредственной ко-окурентности. Ниже представлена таблица Топ-20 биграмм по показателю семантической близости:

| Биграмма (прилагательное, существительное) | Семантический вес |
|---|---|
| программный, обеспечение | 630.23 |
| программный, инженерия | 335.37 |
| первый, очередь | 227,54 |
| заработный, плата | 205.29 |
| интересный, проект | 150.29 |
| точный, наука | 146.28 |
| современный, мир | 127.87 |
| программный, продукт | 120.32 |
| карьерный, лестница | 115.02 |
| карьерный, рост | 111.54 |
| высокий, оплата | 95.92 |
| огромный, количество | 92.31 |
| высокий, уровень | 85.28 |

| | |
|---|---|
| крупный, компания | 82.60 |
| высокий, зарплата | 81.86 |
| собственный, проект | 78.71 |
| новый, язык | 77,58 |
| сложный, задача | 76.01 |
| высокий, плата | 71.19 |
| базовый, знание | 68.53 |

Показатели семантической близости подтверждают статистически значимые связи, выявленные ранее. Например, сочетания «программное обеспечение», «программная инженерия», «заработная плата», «карьерный рост» демонстрируют не только высокую частотность, но и сильную смысловую когезию в текстах студентов, подчеркивая их центральную роль в формировании представлений о профессии.

Дальнейший анализ включал классификацию выявленных биграмм по тематическим категориям, что позволило выявить ключевые аспекты, волнующие студентов:

- **Финансовые аспекты:** (56 уникальных сочетаний, общая частота 226, средний семантический вес 5.24). Отчетливо демонстрирует значимость материального благополучия как мощного мотивационного фактора для студентов, выбирающих IT-сферу. Примеры: "высокий оплата", "высокий труд", "высокий зарплата".
- **Необходимые навыки и знания:** (120 уникальных сочетаний, общая частота 488, средний семантический вес 3.47). Указывает на осознание студентами важности непрерывного обучения и развития, потребности в освоении различных языков программирования и фундаментальных знаний. Примеры: "необходимое программирование", "необходимые знания", "различные языки".
- **Характер работы:** (52 уникальных сочетания, общая частота 191, средний семантический вес 4.63). Свидетельствует о том, что студенты воспринимают работу программиста как деятельность, требующую креативности, предоставляющую возможности для реализации интересных проектов и решения сложных задач. Примеры: "интересный проект", "творческая разработка", "сложная задача", "собственный проект".
- **Востребованность и стабильность работы:** (70 уникальных сочетаний, общая частота 237, средний семантический вес 3.66). Указывает на то, что студенты оценивают профессию программиста как стабильную, востребованную и предоставляющую широкие возможности трудоустройства. Примеры: "реальная работа", "неотъемлемая работа", "крупная компания".
- **Личные качества:** (2 уникальных сочетания, общая частота 4, средний семантический вес 4.61). Несмотря на малочисленность, говорит об

осознании необходимости личной ответственности в профессиональной деятельности и жизни в целом. Примеры: "ответственный шаг", "ответственная жизнь".

Классификация биграмм позволяет сделать вывод, что студенты, выбирающие профессию программиста, в своих эссе акцентируют внимание на финансовой привлекательности, необходимости постоянного развития и углубления знаний, интересном и творческом характере самой работы, высокой востребованности и стабильности на рынке труда, а также, хотя и менее выраженно, на осознании личной ответственности.

# ЗАКЛЮЧЕНИЕ

Настоящая диссертационная работа успешно решила актуальную проблему исследования лексических особенностей дискурса, отражающего представления студентов о профессии программиста. Поставленная цель – разработка и применение метода анализа коллокаций «прилагательное + существительное» в корпусе студенческих эссе – была достигнута в полной мере.

В ходе исследования был сформирован репрезентативный корпус, разработан и реализован программный комплекс для автоматизированной обработки текстов, что позволило эффективно извлечь и проанализировать целевые коллокации. Количественный анализ с использованием статистических мер (PMI, t-критерий, LLR) и последующий качественный (семантический и тематический) анализ коллокаций выявили доминирующие аспекты образа профессии программиста в сознании студентов. Среди них особо выделяются финансовые аспекты, важность навыков, характер работы, востребованность и личные качества.

Научная новизна работы заключается в комплексной методологии корпусного анализа, интегрирующей многомерный статистический подход и тематическую классификацию для изучения профессионального самосознания. Теоретическая значимость состоит в углублении знаний о формировании профессионального языка. Практическая значимость проявляется в возможности применения результатов для совершенствования образовательных программ, профориентационной работы и анализа мотиваций в IT-сфере.

Таким образом, работа не только подтвердила эффективность примененных методов корпусной лингвистики, но и предоставила ценные эмпирические данные, способствующие более глубокому пониманию профессионального самоопределения будущих специалистов.

# СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1-А] «Статистический анализ коллокаций в эссе студентов для создания профессиограммы программиста» / Т.Б. Джаримбетов // Материалы CDV международной научно-практической конференции. – С. 279–284.

[2-А] «Анализ текста методом извлечения коллокаций» / Т.Б. Джаримбетов // Компьютерные системы и сети : сборник материалов 61-й научной конференции аспирантов, магистрантов и студентов БГУИР (Минск, 22–26 апреля 2025 г.). – Минск, 2025. – С. 144–146.