На правах рукописи

Гукасян Цолак Гукасович

МЕТОДЫ И ПРОГРАММНЫЕ СРЕДСТВА ДЛЯ ВЫЯВЛЕНИЯ ЗАИМСТВОВАНИЙ В ТЕКСТАХ НА АРМЯНСКОМ ЯЗЫКЕ

Специальность 05.13.11 —
«Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Автореферат

диссертации на соискание учёной степени

кандидата технических наук

Москва — 2021

Работа выполнена в Российско-Армянском университете.

Научный руководитель: кандидат физико-математических наук
Турдаков Денис Юрьевич

Официальные оппоненты: **Лукашевич Наталья Валентиновна,**
доктор технических наук,
Научно-исследовательский вычислительный центр МГУ имени М.В. Ломоносова,
ведущий научный сотрудник

Чехович Юрий Викторович,
кандидат физико-математических наук,
Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук,
Заведующий отделом

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования ”Казанский (Приволжский) федеральный университет”

Защита состоится 25 мая 2021 г. в 12 часов на заседании диссертационного совета Д 002.087.01 на базе Федеральном государственном бюджетном учреждении науки Институте системного программирования им.В.П.Иванникова РАН по адресу: 109004, г. Москва,ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Институте системного программирования им. В. П. Иванникова РАН.

Автореферат разослан _____ 2021 года.

Ученый секретарь
диссертационного совета
Д 002.087.01, кандидат физ.мат. наук Зеленов С. В.

Общая характеристика работы

Актуальность темы. Определение степени уникальности работ является одной из самых серьезных проблем в научных исследованиях. Неуникальные, заимствованные работы (заимствованием считается как правильно процитированный текст, так и плагиат, что является нарушением авторских прав), которые остаются нераскрытыми, могут иметь серьезные негативные последствия по нескольким причинам.

Заимствованные исследовательские работы препятствует научному процессу, например, искажая механизмы отслеживания и исправления результатов. Если исследователи расширят или пересмотрят более ранние результаты в последующих исследованиях, то статьи, содержащие заимствования из исходной статьи, останутся неизменными. Неправильные результаты могут распространиться и повлиять на последующие исследования или практическое применение. Исследования показывают, что некоторые частично или полностью заимствованные работы цитируются по крайней мере так же часто, как и оригинал. Это проблематично, поскольку число цитирований является широко используемым показателем эффективности исследований, например, для принятия решений о финансировании или найме. Отсутствие надежных механизмов выявления и предотвращения случаев карьерного продвижения путем сплагиаченного труда может привести к кризисным ситуациям в различных отраслях общественной жизни (в образовательной [1] [2] и судебной[3] системах, например). С образовательной точки зрения заимствования наносят ущерб приобретению и оценке компетенций. Было выявлено, что учащиеся армянских вузов в целом осведомлены, какие именно действия считаются плагиатом, но продолжают их совершать из-за отсутствия мер пресечения. Кроме того, заимствованные работы тратят ресурсы. Опыт показывает, что расследования уникальности работа часто требуют сотен рабочих часов от затронутых учреждений, и поэтому очень важно наличие автоматической системы обнаружения заимствований, с помощью которой можно будет сократить затраты.

[1] https://ru.armeniasputnik.am/society/20190821/20137912/Epopeya-s-AGEU-zakonchilas-molodoy-io-rektora-Ruben-Ayrapetyan-pokinet-svoy-post-.html

[2] https://ru.armeniasputnik.am/society/20200525/23163731/Esli-moy-zam-pokryvaet-plagiat-dissertatsii-on-dolzhen-otvetit—-Araik-Arutyunyan-.html

[3] https://news.am/rus/news/514896.html

Быстрое развитие информационных технологий, особенно Интернета, сделало заимствование работ легче, чем когда-либо. В 2015 году было проведено исследование образовательной политики Армении в направлении усиления академической добросовестности, которое подтвердило, что незаконные заимствования в курсовых, бакалаврских и магистерских работах являются одним из самых распространенных нарушений. Заимствованием, кроме дословного копирования, считается сокрытие заимствований путем перефразирования и перевода. Уникальность работы искусственно увеличивают также с помощью технических приемов, которые используют слабые места методов извлечения текста системы обнаружения заимствований и меняют исходный документ таким образом, чтобы его текст визуально не менялся, но доля обнаруженных заимствований получалась маленькой.

Исследование и разработка методов выявления заимствований сейчас является довольно популярной, если судить по количеству опубликованных статей в последние годы. Тем не менее, для многих языков не существует специализированной системы обнаружения заимствований. В таких случаях приходится прибегать к использованию инструментов, не адаптированных к определенному языку, однако эти решения как правило не учитывают особенности языка и не показывают достаточный уровень качества обработки. В существующих работах, посвященных обнаружению заимствований для армянского языка, не реализованы (или не работают без ручного вмешательства пользователя) обнаружение парафраза, случаев замены омоглифов и других технических приемов маскировки заимствований, поиск релевантных источников в Интернете.

Учитывая недостаток исследований и решений в этой области для армянского языка, адаптация и разработка методов выявления заимствований в армянских текстах очень актуальна. В настоящее время, армянские университеты и другие заинтересованные организации вынуждены либо отказываться от проверки текстов на уникальность, либо использовать универсальные инструменты, которые не адаптированы к армянскому языку и, как правило, способны находить только случаи полного дублирования. Например, Российско-Армянский университет в Ереване использует систему Антиплагиат.ру [4], однако изучение этой системы показало, что она неспособна выявлять замену синонимов, парафраз, использование технических приемов для армянского языка.

[4] https://www.antiplagiat.ru

Выявление перечисленных видов заимствований требует использование таких инструментов, как языковые модели, оптическое распознавание текста и т.д., предназначенных специально для армянского языка.

В этой диссертации рассматриваются методы и системы обнаружения заимствований. С технической точки зрения в литературе различают два подхода к обнаружению заимствований: внешние и внутренние. Внешние методы сравнивают текст проверяемого документа с проверочным набором потенциальных источников. Когда источник заимствований отсутствует в проверочном наборе или когда нет проверочной базы, поиск заимствований производится с помощью внутренних методов, которые основываются исключительно на имеющемся документе для нахождения подозрительных участков.

Целью данной работы является исследование и разработка методов и программных инструментов установления степени уникальности текстов для армянского языка. Объектом исследования данной диссертации являются тексты на литературном армянском языке, в частности, курсовые, выпускные квалификационные работы, диссертации. Предмет исследования – уникальность текстов. В рамках этой работы степень уникальности определяется как процент текста, не встречающийся в других работах. В диссертации не изучаются политические и юридические аспекты, связанные с заимствованием ранее опубликованных работ. Также не учитывается отношение автора к этому действию.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. На основе анализа существующих решений разработать и реализовать внутренние стилометрические методы нахождения заимствований для армянского языка.
2. На основе анализа существующих решений разработать и реализовать методы борьбы с техническими методами маскировки заимствований для армянского языка.
3. На основе анализа существующих решений разработать и реализовать метод нахождения нечетких дубликатов для армянского языка.
4. На основе анализа существующих решений разработать и реализовать метод определения парафраза в армянских текстах для внешних методов нахождения заимствований.

5. На основе реализованных методов, создать программную систему для оценки степени уникальности текстовых документов на армянском языке.

Для достижения поставленной цели и решения вышеуказанных задач были изучены и применены методы машинного обучения, математической статистики, информационного поиска нечетких дубликатов, математического анализа и линейной алгебры. Для программной реализации алгоритмов и разработки системы использовались методы объектно-ориентированного программирования.

Основные положения, выносимые на защиту:

1. Предложен подход к извлечению векторных представлений слов, полностью основанных на признаках на уровне подслов (символов и морфем), который для языков с богатой морфологией позволяет смягчить проблему разреженности данных и сокращает количество параметров в модели. На основе таких векторных представленияй слов получены модели лемматизации и морфологического анализа текстов, требующие гораздо меньше памяти и при этом не уступающие в точности моделям, имеющим в несколько раз больше параметров.
2. Разработаны методы и программные инструменты для автоматизации процесса построения размеченных наборов данных для языков с ограниченными ресурсами. На основе предложенных и существующих подходов для армянского языка впервые созданы размеченные наборы текстов для задач распознавания именованных сущностей, обнаружения парафраза, векторного представления слов, стилометрического анализа, и исправления ошибок автоматического распознавания текстов. Для первых трех из перечисленных задач созданы тестовые наборы с ручной разметкой. Разработанные наборы данных позволили создать программные инструменты для соответствующих задач, превосходящие по точности существующие аналоги.
3. С использованием перечисленных выше инструментов разработана и внедрена программная система для оценки степени уникальности текстовых документов на армянском языке, которая позволяет обнаружить полное и частичное дублирование, парафраз, техническую

маскировку, а также выполняет поиск заимствований как в проверочной базе документов, так и в Интернете.

Научная новизна: Представлен новый метод генерации парафразов предложения с помощью обратного машинного перевода в несколько итераций и ручной проверки корректности результатов перевода[1]. Разрабатан метод автоматической генерации обучающих и тестовых примеров для задач нахождения и исправления ошибок оптического распознавания текстов [2]. Представен новый подход к использованию Викиданных и статей Википедии для полной автоматизации процесса генерации обучающих примеров для задачи распознавания именованных сущностей [3].

Предложены модификации модели векторов fastText на основе подслов, которые решают проблему разреженности данных для языков с богатой морфологией и существенно сокращают размер этих моделей без серьезной потери точности в задачах леммatизациии морфологического анализа [4; 5].

Практическая значимость Основная практическая значимость диссертации заключается в разработанной системе оценки уникальности текстов, которая может быть применена в работе высших учебных заведений, ВАК РА и других похожих организаций. Для армянского языка впервые разработаны программные инструменты, позволяющие выполнять внутренний стилометрический анализ текстов на наличие заимствований, обнаруживать парафраз, исправлять ошибки в результатах оптического распознавания текстов.

Впервые для армянского языка созданы тестовые наборы данных с ручной разметкой для задач распознавания именованных сущностей [3], определения парафраза [1], внутреннего стилометрического анализа текстов [6], а также оценки качества векторных представлений слов [7]. Созданные размеченные наборы текстов могут быть использованы в будущих исследованиях для разработки и оценки качества инструментов обработки армянских текстов.

Предложенные автоматические методы генерации размеченных данных позволяют сократить использование человеческих и других ресурсов при создании обучающих и тестовых данных для соответствующих задач, могут быть применены для создания размеченных наборов текстов для других языков.

Апробация работы. Результаты данной работы докладывались на конференциях, форумах:

1. Science and Technology Convergence (STC) Forum 2018, Ереван, РА;

2. Открытая конференция ИСП РАН им. В.П. Иванникова 2018, Москва, РФ;
3. XIV Годичная научная конференция РАУ, 2019, Ереван, РА;
4. Международная конференция ”Иванниковские чтения 2020, Орел, РФ;

Публикации. По теме диссертации опубликовано 7 печатных работ, в том числе три статьи [1; 3; 5] в изданиях и сборниках научных конференций, индексируемых в Scopus, две статьи [4; 6] в рецензируемых научных журналах из перечня ВАК РФ по специальности 05.13.11, и две статьи [2; 7] в других изданиях.

Личный вклад. Предлагаемые в диссертации инструменты, текстовые наборы данных и исследования разработаны и выполнены автором или при его непосредственном участии. Автор имеет решающий вклад в планировании совместных работ [1—3; 5—7], разработке и адаптации методов автоматической разметки и обработки текстов, принимал непосредственное участие в сборе, подготовке и ручной разметке текстов, планировании и проведении экспериментов. В публикации [5] автору принадлежит основная часть, совместно проводилось измерение качества разработанных моделей.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава посвящена описанию используемой типологии заимствований. Так как целью данной работы является исследование и разработка методов установления степени уникальности текстов для армянского языка, необходимо определить какие именно участки текста считаются неуникальными, то есть заимствованием.

В этой диссертации рассматриваются методы обнаружения следующих форм заимствований: буквальные заимствования, техническая маскировка, замена синонимов и парафраз. Исследование методов обнаружения перевода оставляется на будущее.

В работе не изучаются политические и юридические аспекты, связанные с заимствованием ранее опубликованных работ, а также отношение автора к этому действию. Поэтому в применяемом определении неуникального фрагмента текста, включены все перечисленные виды заимствований:

1. Незаконные заимствования;
2. Переиспользование своих более ранних работ;
3. Заимствование с некорректным указанием источника (например, когда студент не знает, как правильно цитировать источники в конкретном стиле);
4. Законные заимствования (с согласия автора оригинала).

Вторая глава посвящена исследованию изучению и разработке внутренних методов обнаружения заимствований, в частности стилометрического анализа и выявления технических методов маскировки.

Для стилометрического анализа были проведено исследование подходов для разных постановок задач внутреннего анализа: обнаружение изменения стиля в документе (style change detection), обнаружение границ нарушений стиля (style breach detection), кластеризация по авторству (author clustering). На основе проведенного анализа для армянского языка была выбраны и реализованы модели (Nath et al. 2019), (Karaś et al. 2017), а также метод обнаружения границ нарушений стиля на основе иерархической кластеризации и собственного набора стилометрических признаков. Для признакового описания стиля написания текста были скомпилированы лингвистические ресурсы (списки аббревиатур, редких, жаргонных слов), которые в дальнейшем также могут быть использованы в разработке других инструментов обработки армянских текстов. Используя тематически и по смыслу близкие тексты, для тестирования моделей автоматически были сгенерированы примеры из учебников, диссертаций, энциклопедий, новостных и художественных текстов. Экспериментами было установлено, что реализованные стилометрические методы работают лучше, чем случайные базовые решения (Табл. 1, Рис. 1), но тем не менее не показывают достаточно высокий уровень точности для применения на практике. Результаты этого исследования опубликованы в статье [6].

Также были исследованы распространенные методы технической маскировки заимствований и искусственного увеличения степени уникальности работы и методы их обнаружения. На основе существующих подходов для других

Жанр текстов		Точность	Полнота	F1	Специфичность	FPR
Академический	Диссертации	0.9002	0.8105	0.853	0.3432	0.6567
	Учебники	0.9217	0.8217	0.8688	0.4	0.6
	Энциклопедии	0.4074	0.55	0.468	0.3333	0.6666
Художественный		0.8437	0.675	0.75	0.5	0.5
Новости		0.0417	0.1428	0.0645	0.7012	0.2987

Таблица 1 — Качество обнаружения изменения стиля модели Nath et al.

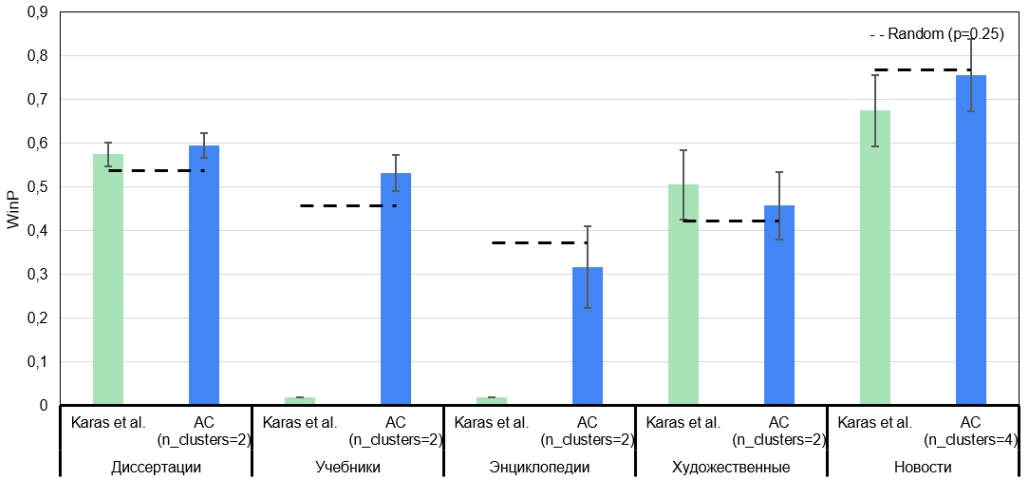

Рис. 1 — Сравнение наиболее точных (90% доверительный интервал) моделей обнаружения границ нарушений стиля и случайного классификатора для каждого жанра.

языков, адаптированы и предложены механизмы и средства выявления скрытого текста, вставки изображений, и замены омоглифов в текстах на армянском языке.

Например, в армянском языке чтобы скрыть незаконное заимствование, злоумышленник может во всех словах заменить все символы ‘o’ (U+0585), ‘h’ (U+0570) на их латинские омоглифы ‘o’ (U+006F), ‘h’ (U+0068). Слова «հայերեն» и «հայերեն» кажутся идентичными человеку, однако во втором слове армянский символ U+0570 заменен латинским U+0068, и если система обнаружения плагиата не предназначена для работы с такими изменениями, то этот трюк может быть использован для скрытия заимствований в тексте. Помимо единичных символов, Юникод также содержит 5 лигатур армянского языка: ﬓ, ﬔ, ﬕ, ﬖ, ﬗ, которые могут быть использованы для замены отдельных глифов соответствующих символов. Перечисленные лигатуры редко встречаются в современной литературе, поэтому злоумышленник может потенциально использовать их для искажения текста и усложнения его обработки системой выявления заимствований.

Для армянского языка списки омоглифов доступны через инструменты homoglyphs[5] и confusable_homoglyphs[6]. Используя эти списки, можно искать замену омоглифов в текстах на армянском языке: если в слове встречается символ, принадлежащий другому алфавиту, то он отмечается как попытка маскировки плагиата в случае, когда этот символ имеет омоглиф в армянском алфавите и при его замене армянским аналогом получится слово, присутствующее в словаре. Эти методы внутреннего анализа, вместе со стилометрическими, были реализованы в программной системе по обнаружению заимствований в текстах.

Третья глава посвящена исследованию внешних методов обнаружения заимствований. Были изучены глобальные и локальные методы обнаружения нечетких дубликатов текстов, в частности алгоритмы поиска нечетких дубликатов документов в проверочной коллекции, методы нахождения полных и нечетких дубликатов на уровне небольших участков текста (коэффициенты Жаккара и Шимкевича-Симпсона, метод отпечатков, модели обнаружения парафраза).

Для применения в системе обнаружения заимствований был выбран метод шинглов, реализация которого с помощью хеширования позволяет выполнять быстрый поиск схожих текстов. Также были изучены подходы к поиску источников заимствований в Интернете, в частности, решения из серии соревнований PAN по информационному поиску. Был выбран алгоритм (Prakash et al. 2014), как наиболее эффективный алгоритм достижения приемлемого уровня полноты результатов при умеренном количестве запросов к поисковой системе, который был адаптирован и реализован в системе поиска заимствований в текстах на армянском языке.

Для детального анализа текстов в системе обнаружения заимствований были изучены модели обнаружения парафраза, методы на основе отпечатков, коэффициентов Жаккара и Шимкевича-Симпсона. Предложен полуавтоматический подход к генерации парафразов предложений на основе обратного перевода (рис. 2). Используя этот подход, для армянского языка впервые был разработан набор парафразов с высоким уровнем лексического разнообразия (табл. 3 и 2). Изученные методы детального анализа были адаптированы и реализованы в программной системе. Используя предобученную нейросетевую модель M-BERT, для армянского языка впервые был создан программный инструмент обнаружения парафраза путем дообученения M-BERT на сгенерированном на-

[5]https://pypi.org/project/homoglyphs/

[6]https://pypi.org/project/confusable_homoglyphs/

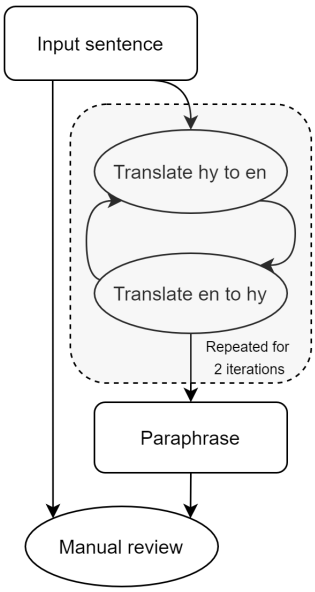

Рис. 2 — Схема генерации парафраза путем перевода из армянского (hy) в английский (en) и обратно.

боре парафразов. Результаты экспериментов по оценке качества полученных моделей определения парафраза представлены в табл. 4. Данное исследование опубликовано в статье [1].

Набор данных	Уровень разнообразия парафразов	
	Обучающие примеры	Тестовые примеры
MRPC	6.79	7.01
ParaPhraser.ru	5.02	5.51
ARPA	8.70	8.66

Таблица 2 — Уровень разнообразия парафразов в корпусах для английского, русского и армянского языков.

В **четвертой главе** приведено описание вспомогательных методов обработки текстов, в частности описаны:

- Методы лемматизации, которые используются для нормализации текста во время индексации, стилометрического анализа и на этапе детального анализа. Дополнительно для языков с богатой морфологией предлагаются модификации стандартных подходов, использующие только символы и морфемы слов для вычисления вектора слова;
- Методы исправления ошибок автоматического распознавания текста;
- Методы распознавания именованных сущностей, используемые для признакового описания текстов в стилометрическом анализе;

Набор данных	Парафразы		Непарафразы		Общее число примеров
	Кол-во примеров	Средний коэффициент Жаккара	Кол-во примеров	Средний коэффициент Жаккара	
Тестовый набор					
MRPC	1147	0.438	578	0.322	1725
ParaPhraser.ru	1137	0.317	762	0.169	1899
ARPA	1021	0.327	661	0.172	1682
Обучающий набор					
MRPC	2753	0.444	1323	0.325	4076
ParaPhraser.ru	4255	0.306	2947	0.119	7202
ARPA	1339	0.320	2894	0.056	4233

Таблица 3 — Распределение парафразов и непарафразов в корпусах для английского, русского и армянского языков.

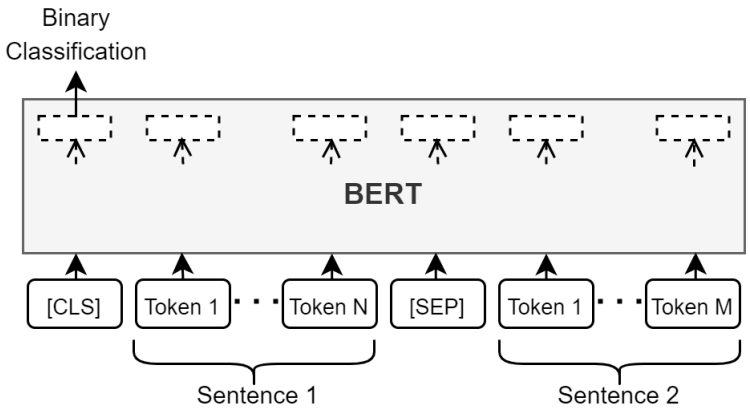

Рис. 3 — Архитектура модели обнаружения парафраза.

Модель	Оценка качества (95% доверительный интервал)			
	F1	Accuracy	Полнота	Точность
a.i. tr-MRPC	0.801 ± 0.0141	0.699 ± 0.0283	0.993 ± 0.0051	0.672 ± 0.0212
a.ii. tr-ParaPhraser	0.838 ± 0.0018	0.771 ± 0.0021	0.977 ± 0.0054	0.734 ± 0.0016
a.iii. ARPA	0.837 ± 0.0028	0.775 ± 0.0028	0.952 ± 0.0089	0.747 ± 0.0023
a.iv. Combined	0.840 ± 0.0017	0.776 ± 0.0017	0.971 ± 0.0056	0.741 ± 0.0013
b. RUBERT	0.837	0.764	0.998	0.721
c. BERT	0.779	0.656	1.0	0.638

Таблица 4 — Результаты оценки качества моделей обнаружения парафраза на тестовом наборе ARPA.

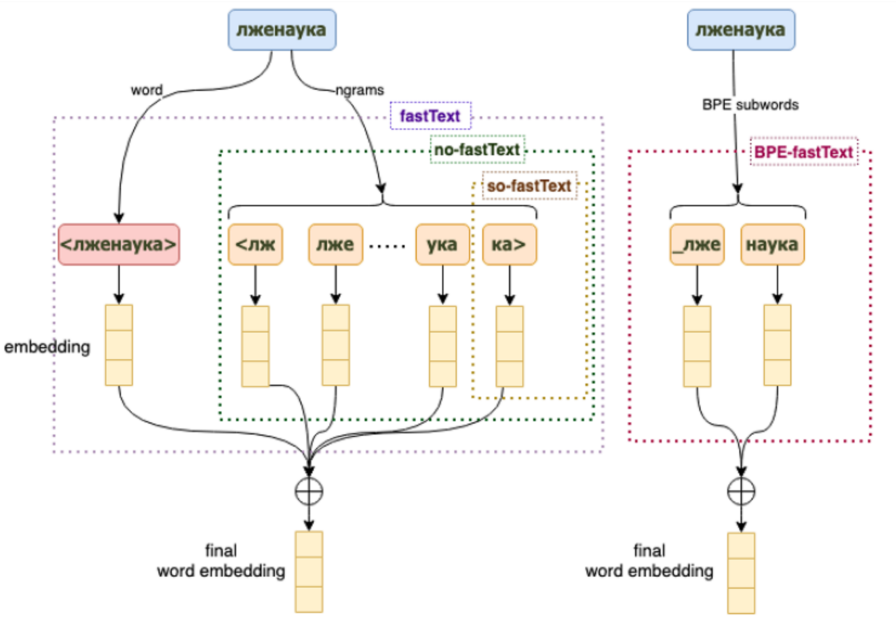

Рис. 4 — Архитектура исходной модели fastText и предлагаемых модификаций.

– Способы обучения и оценки моделей векторного представления слов армянского языка, используемых в качестве признаков в задачах обработки текстов.

Лемматизация В диссертации разработаны модели к лемматизации армянских текстов. На основе анализа существующих методов, основанных на глубоком обучении и показывающих точные результаты на языках с богатой морфологией и маленьким обучающим корпусом, в качестве базы для лемматизатора выбрана нейронная сеть COMBO. Для замены тяжелых по памяти и количеству параметров предобученных моделей векторов слов, предложены несколько альтернатив на основе подслов. Разработаны и протестированы модификации алгоритма векторного представления слов fastText, использующие только внутренние символьные N-граммы, суффиксы и морфемы слова (рис. 4). В качестве морфем слова используются его подслова, получаемые в результате BPE кодирования. Последнее позволяет уменьшить размер модели лемматизации от 1 Гб до нескольких Мб, а размер предобученных моделей векторов уменьшается примерно в 100 раз, при этом без потери точности. Эти модификации особенно актуальны для языков с богатой морфологией, так как позволяют смягчить проблему разреженности данных и сократить количество параметров в модели. Результаты этого исследования опубликованы в статьях [4; 5].

Разработанный лемматизатор превосходит существующие аналоги, достигая точности, сравнимой с state-of-the-art для малоресурсных языков.

Модель векторов	Accuracy	Accuracy на OOV	Кол-во параметров векторов	Общее кол-во параметров
fastText	90.77	75.25	1.2×10^8	1.3×10^8
no-fastText	90.36	73.96	4.0×10^7	5.0×10^7
so-fastText	89.55	72.31	4.0×10^7	5.0×10^7
BPEmb	90.61	74.92	2.5×10^6	1.2×10^7
BPE-custom	90.86	75.62	1.2×10^6	1.1×10^7

Таблица 5 — Сравнение качества (Accuracy) лемматизаторов на основе разных моделей векторного представления слов.

Модели и наборы данных для распознавания и классификации именованных сущностей Разработаны методы и программные инструменты для автоматизации процесса построения наборов данных для языков с ограниченными ресурсами. На основе предложенного подхода создан не имевший аналогов эталонный корпус армянского языка для задачи распознавания именованных сущностей.

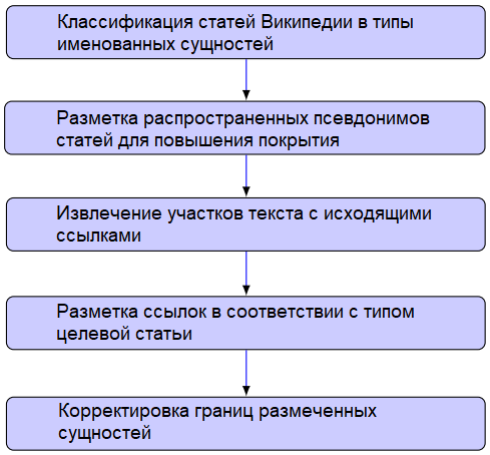

Рис. 5 — Алгоритм автоматической генерации размеченных предложений.

Для решения задачи распознавания именованных сущностей было проведено исследование были разработаны наборы размеченных данных серебряного и золотого стандартов, а также установлены базовые результаты на популярных моделях. В результате был создан корпус именованных сущностей с 163000 токенами, автоматически сгенерированный из Википедии, и еще один корпус новостных предложений с 53400 токенами с ручной разметкой именованных сущностей: людей, организаций и географических объектов. Использованный подход может в дальнейшем использоваться для автоматичекой гене-

Значение атрибута subclass of	Тип сущности
company, business enterprise, company, juridical person, air carrier, political organization, government organization, secret service, political party, international organization, alliance, armed organization, higher education institution, educational institution, university, educational organization, school, fictional magic school, broadcaster, newspaper, periodical literature, religious organization, football club, sports team, musical ensemble, music organisation, vocal-musical ensemble, sports organization, criminal organization, museum of culture, scientific organisation, non-governmental organization, nonprofit organization, national sports team, legal person, scholarly publication, academic journal, association, band, sports club, institution, medical facility	ORG
state, disputed territory, country, occupied territory, political territorial entity, city, town, village, rural area, rural settlement, urban-type settlement, geographical object, geographic location, geographic region, community, administrative territorial entity, former administrative territorial entity, human settlement, county, province, federated state, district, county-equivalent, municipal formation, raion, nahiyah, mintaqah, muhafazah, realm, principality, historical country, watercourse, lake, sea, still waters, body of water, landmass, minor planet, landform, natural geographic object, mountain range, mountain, protected area, national park, geographic region, geographic location, arena, bridge, airport, stadium, performing arts center, public building, venue, sports venue, church, temple, place of worship, retail building	LOC
person, fictional character, fictional humanoid, human who may be fictional, given name, fictional human, magician in fantasy	PER

Таблица 6 — Таблица значений атрибутов subclass of и соответствующей метки типа именованной сущности.

рации корпусов для других языков. Важность тестового корпуса состоит в том, что он может служить эталоном для будущих систем распознавания именованных сущностей, разработанных для армянского языка. Кроме того, чтобы установить применимость основанных на Википедии подходов к армянскому языку, предоставлены результаты оценки для 3 различных систем распознавания именованных сущностей (табл. 7), обученных и протестированных на наших наборах данных. Исследование опубликовано в статье [3].

Модель	Dev			Test		
	Точность	Полнота	F1	Точность	Полнота	F1
Stanford NER	76.86	70.62	73.61	**78.46**	46.52	58.41
spaCy 2.0	68.19	71.86	69.98	64.83	**55.77**	59.96
Char-biLSTM+biLSTM+CRF	**77.21**	**74.81**	**75.99**	73.27	54.14	**62.23**

Таблица 7 — Оценка качества распознавания алгоритмов.

Исправление ошибок автоматического распознавания. Рассмотрена проблема автоматического распознавания текста на армянском языке, в частности, решение задачи исправления ошибок автоматического распознавания. Результаты были опубликованы в статье [2]. Используется двухэтапный подход к решению задачи: (i) обнаружение ошибок распознавания с помощью многослойного персептрона и (ii) исправление ошибок с помощью преобразователя последовательности на основе сверточной нейронной сети.

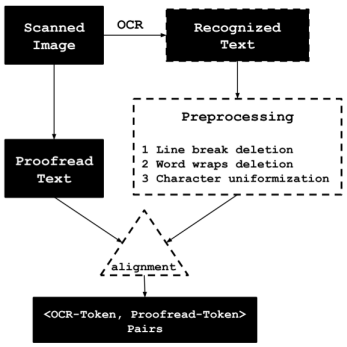

Рис. 6 — Схема генерации примеров для обучения и тестирования.

Чтобы сгенерировать наборы данных для обучения и оценки качества моделей, к отсканированным изображениям текстов был применен OCR, вывод которого был согласован с исправленными версиями (рисунок 6).

Разработанные инструменты постобработки позволили уменьшить коэффициент ошибок в словах (WER) из статей Советско-Армянской Энциклопедии, распознанных Tesseract OCR, на 23.5%, достигая результатов, сравнимых с коммерческими решениями данной задачи (табл. 8). Результаты опубликованы в [2].

Модель	WER	Доступность
Google Docs	1.10	Бесплатный
Convertio	0.27	Коммерческий
ABBYY FineReader	0.54	Коммерческий
Tesseract	0.51	Бесплатный
Tesseract + постобработка	0.39	Бесплатный

Таблица 8 — Сравнение качества Tesseract с исправлением ошибок и результатов других инструментов.

Обучение и оценка качества предобученных векторных представлений Также была выполнена оценка предобученных векторных представлений слов армянского языка в 3 различных задачах. С этой целью были созданы эталонные наборы данных для внутренней и внешней оценки представлений слов. Для внутренних тестов слово была переведена и адаптирована «задача аналогий» (табл. 9). Результаты оценки приведены в табл. 10.

Для внешней оценки используется задача морфологического анализа (табл. 11), а также создается корпус новостных текстов для сравнения качества в задаче классификации (12). Дополнительно, была собрана коллекция неразмеченных текстов, на которой были обучены новые более эффективные модели

Раздел	Вопросы	Тип примеров
capital-common-countries	506	Семантический
capital-world	4369	Семантический
currency	866	Семантический
city-in-state	56	Семантический
family	506	Семантический
gram1-adjective-to-adverb	992	Синтаксический
gram2-opposite	812	Синтаксический
gram3-superlative	1122	Синтаксический
gram4-present-participle	1056	Синтаксический
gram5-nationality-adjective	1599	Синтаксический
gram6-past-tense	1560	Синтаксический
gram7-plural	1332	Синтаксический
gram8-plural-verbs	870	Синтаксический

Таблица 9 — Разделы адаптированной задачи аналогий слов.

Модель	capital-common	capital-world	curr.	city-state	family	gram1	gram2	gram3	gram4	gram5	gram6	gram7	gram8
fastText[Wiki,.text]	5.34	0.78	0	0	4.54	14.91	30.29	39.57	7.19	44.21	23.71	29.35	0.45
fastText[Wiki,.bin]	5.34	0.77	0	0	4.54	16.53	30.29	27.98	7.19	47.52	23.71	29.35	0.45
fastText[CC,.text]	32.61	11.42	2.77	7.14	13.83	22.07	30.66	43.76	4.45	41.58	18.33	19.96	5.51
fastText[CC,.bin]	72.53	39.28	11.55	48.21	47.83	25.2	36.21	49.64	19.7	8.53	23.08	41.89	41.03
fastText[new,.text]	27.66	8.1	0.1	1.79	16.2	28.02	41.74	48.3	23.95	54.59	50.51	53.67	6.09
fastText[new,.bin]	27.67	8.1	1.03	1.79	16.2	30.14	41.74	58.82	23.95	54.59	50.51	53.67	6.09
SkipGram[YerevaNN]	39.32	10.66	2.07	8.93	5.73	4.03	0.61	3.74	1.23	23.57	0.12	5.78	0.8
SkipGram[new]	36.17	17.37	2.3	3.57	17.79	7.56	12.43	16.39	1.7	37.77	4.93	16.81	10.34
CBOW[new]	28.65	13.04	1.5	5.36	29.05	10.48	14.77	17.91	5.49	24.26	6.98	28.6	11.83
GloVe[pioNER d50]	8.1	1.06	0.11	0	6.71	3.32	2.46	3.29	0.94	11.75	1.02	6.98	1.26
GloVe[pioNER,d100]	10.67	1.67	0.46	3.57	10.27	4.53	5.41	4.72	1.51	16.51	1.15	7.43	3.9
GloVe[pioNER,d200]	10.67	2.28	0.8	7.14	11.66	3.52	8.74	7.13	1.32	15.69	1.02	5.7	2.87
GloVe[pioNER,d300]	10.87	2.05	0.46	5.36	11.46	3.22	7.88	5.79	0.94	13.75	0.7	4.72	1.49
GloVe[new]	75.3	49.14	2.19	23.21	15.8	6.55	11.2	12.74	2.27	47.71	1.85	20.49	5.4

Таблица 10 — Точность (Accuracy, %) векторных представлений слов на разделах адаптированной задачи аналогий.

векторов GloVe, CBOW, SkipGram, fastText. Исследование было опубликовано в [7].

Модель	Валидационный набор		Тестовый набор	
	UPOS	FEATS	UPOS	FEATS
fastText[Wiki,.text]	92.99	83.83	89.54	81.03
fastText[Wiki,.bin]	89.27	75.96	84.35	71.14
fastText[CC,.text]	91.86	79.85	88.38	75.44
fastText[CC,.bin]	91.54	77.78	87.59	73.29
fastText[new,.text]	94.64	85.89	93.35	83.99
fastText[new,.bin]	91.43	80.1	87.55	74.78
SkipGram[YerevaNN]	91.44	80.04	87.45	75.68
SkipGram[new]	93.9	86.45	91.6	83.77
CBOW[new]	93.87	85.39	92.72	84.07
GloVe[pioNER d50]	91.78	82.42	89.12	80.18
GloVe[pioNER,d100]	91.78	82.95	89.06	80.51
GloVe[pioNER,d200]	92.34	83.77	88.90	80.07
GloVe[pioNER,d300]	91.70	83.19	89.09	80.78
GloVe[new]	94.09	86.23	92.98	83.97

Таблица 11 — Точность (Accuracy, %) морфологического анализа на основе разных моделей векторного представления.

Модель	Accuracy	Точность	Полнота	F1
fastText[Wiki,.text]	66.63	60.38	63.02	58.96
fastText[Wiki,.bin]	65.79	59.78	63.97	58.15
fastText[CC,.text]	65.75	59.15	63.68	57.03
fastText[CC,.bin]	65.51	58.96	63.43	56.94
fastText[new,.text]	63.34	56.97	59.81	55.18
fastText[new,.bin]	60.12	53.66	55.4	51.96
SkipGram[YerevaNN]	64.34	57.87	59.73	56.28
SkipGram[new]	66.68	60.84	63.56	59.83
CBOW[new]	67.92	61.94	65.2	60.94
GloVe[pioNER d50]	64.26	57.57	58.39	54.4
GloVe[pioNER,d100]	65.91	60.15	62.34	58.91
GloVe[pioNER,d200]	68.16	62.13	65.54	60.6
GloVe[pioNER,d300]	67.85	61.7	65.36	60.43
GloVe[new]	69.77	63.93	66.55	63.13

Таблица 12 — Точность классификации текстов на основе разных моделей векторного представления.

В **пятой главе** приведено описание программной системы для обнаружения текстовых заимствований. Сначала приводится краткий обзор существующих систем. Затем описывается общая архитектура системы, его модули и компоненты. В третьем разделе приводится описание метода и программных средств для полнотекстового поиска. Четвертый раздел посвящен технологиям поиска заимствований в сети с помощью поисковых систем, пятый - инструментам и механизму извлечения текста из документов, а шестой раздел предоставляет описание используемых методов реализации асинхронности для вычисления трудоемких задач.

В рамках этой работы была разработана система выявления заимствований (далее - Система), используя микросервисную архитектуру (Рис. 7). В ка-

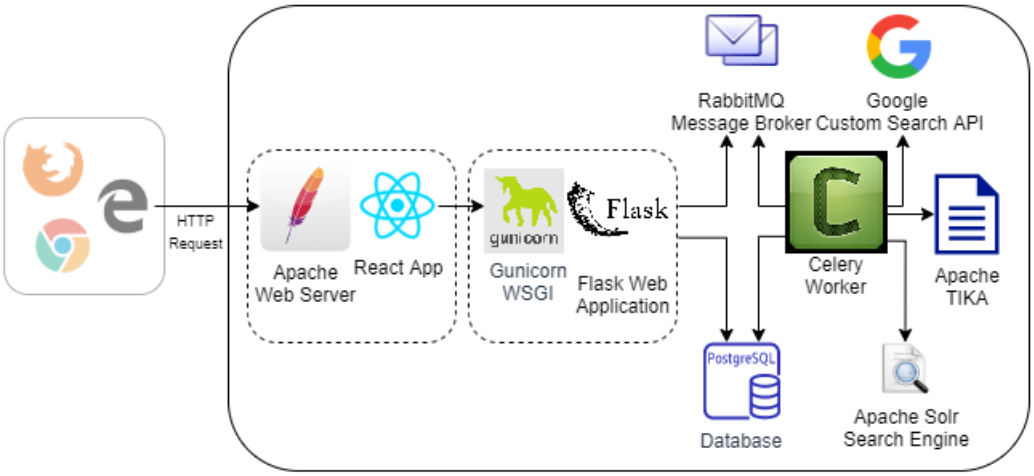

Рис. 7 — Архитектура микросервисов.

честве отдельных микросервисов были организованы веб-сервер Apache с приложением React, основной код бэкенда вместе с WSGI интерфейсом, база данных, модуль полнотекстового поиска (Apache Solr), модуль извлечения текста из документов (Apache TIKA), брокер сообщений RabbitMQ, и задачи Celery. Разработанная система запускается как многоконтейнерное приложение. Для его описания используется пакетный менеджер Docker Compose.

На рисунке 8 изображены основные этапы обработки документа при его проверке. После автоматического распознавания текста документа, в нем исправляются ошибки, далее исправленный текст проходит проверку на наличие технических приемов для маскировки плагиата. Перед применением стилометрических и внешних методов поиска заимствований текст разбивается на фрагменты (параграфы и, опционально, предложения). Для внутреннего анализа применяется метод на основе кластеризации. Внешний анализ сравнивает фрагменты документа с потенциальными источниками из проверочной базы (далее - База). Изначально База состоит из документов, добавленных администратором Системы. Во время проверки документа из его фрагментов извлекаются ключевые словосочетания, которые затем отправляются в систему поиска в Интернете, результаты которого фильтруются и добавляются в Базу. Для поиска в Интернет используется реализация алгоритма Пракаща и др. После завершения поиска в Интернете, каждый фрагмент проверяемого документа целиком отправляется в качестве запроса в систему полнотекстового поиска по документам Базы, которая возвращает список наиболее релевантных фрагментов. Проверяемый фрагмент попарно сравнивается с фрагментами из списка,

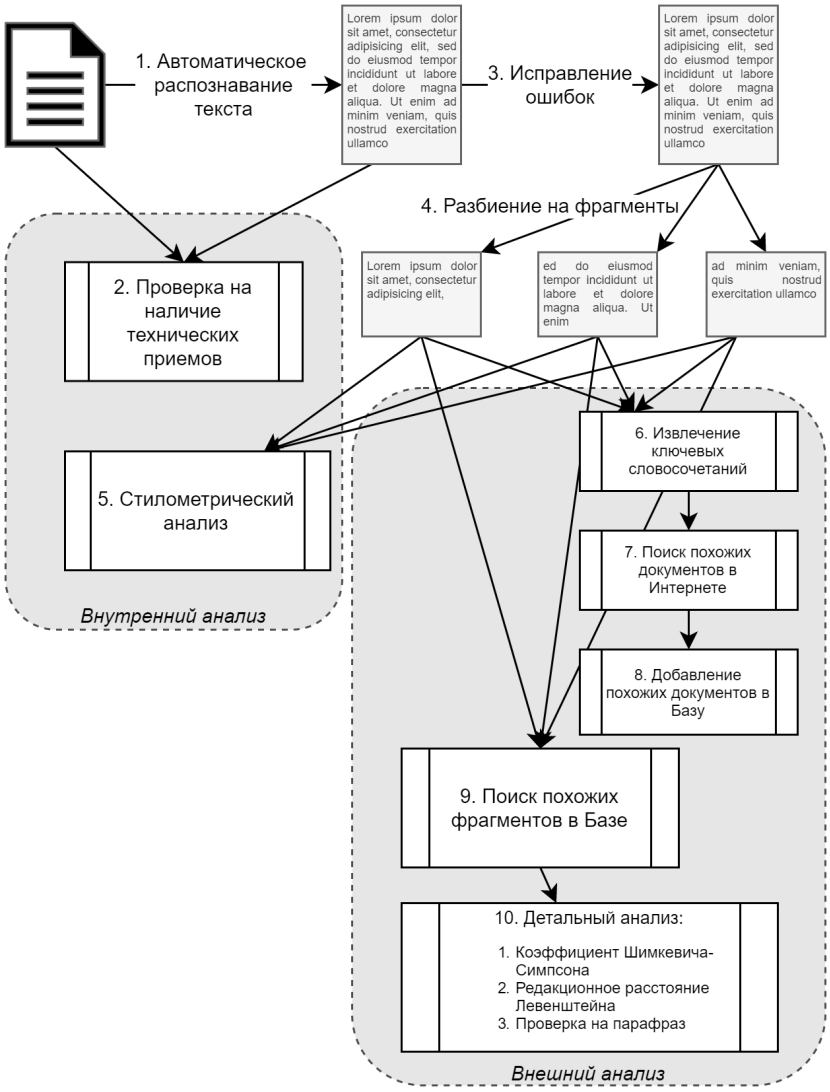

Рис. 8 — Схема использования методов обработки текста в реализованной системе.

используя методы детального анализа, после чего отфильтрованные фрагменты возвращаются в качестве потенциальных источников заимствования. Для реализации полнотекстового поиска используется метод шинглов, для детального анализа фрагменто - пороговые классификаторы на основе коэффициента Шимкевича-Симпсона и редакционное расстояние Левенштейна, и нейросетевой метод обнаружения парафраза.

В **заключении** сформулированы основные результаты работы:

1. Разработаны методы и программные инструменты для автоматизации процесса построения наборов данных для языков с ограниченными ресурсами. На основе предложенных подходов и с помощью ручной разметки впервые для армянского языка созданы размеченные текстовые наборы для задач распознавания именованных сущностей, обнаружения парафраза, векторного представления слов, стилометрического анализа, и исправления ошибок автоматического распознавания текстов. Разработанные наборы позволили создать программные инструменты для соответствующих задач, превышающие по показателям точности существующие аналоги.
2. Предложен подход к извлечению векторных представлений слов, полностью основанных на признаках на уровне подслов (символов и морфем), который для языков с богатой морфологией позволяет смягчить проблему разреженности данных и сокращает количество параметров в модели. На основе таких векторных представленияй слов получены модели лемматизации и морфологического анализа текстов, которые требуют гораздо меньше памяти и не уступают в точности моделям, имеющим в несколько раз больше параметров.
3. С использованием перечисленных выше инструментов разработана и внедрена программная система для оценки степени уникальности текстовых документов на армянском языке, которая помимо прямого копирования позволяет обнаружить нечеткие дубликаты, парафраз, техническую маскировку, выполняет поиск заимствований как в проверочной базе документов, так и в Интернете.

Список литературы

1. *Malajyan A.*, *Avetisyan K.*, *Ghukasyan T.* ARPA: Armenian Paraphrase Detection Corpus and Models // 2020 Ivannikov Memorial Workshop (IVMEM). — 2020. — С. 35—39. — DOI: 10.1109/IVMEM51402.2020.00012.
2. *Tigranyan S.*, *Ghukasyan T.* Post-OCR Correction of Armenian Texts Using Neural Networks. // “Vestnik” Scientific Journal of Russian-Armenian University. — 2020.
3. pioNER: Datasets and Baselines for Armenian Named Entity Recognition / T. Ghukasyan, G. Davtyan, K. Avetisyan, I. Andrianov // 2018 Ivannikov Ispras Open Conference (ISPRAS). — 2018. — С. 56—61. — DOI: 10.1109/ISPRAS.2018.00015.
4. *ГУКАСЯН Ц.* Векторные модели на основе символьных н-грамм для морфологического анализа текстов. // Труды Института системного программирования РАН. — 2020. — Т. 32, № 2. — С. 7—14. — DOI: https://doi.org/10.15514/ISPRAS-2020-32(2)-1.
5. *Ghukasyan T.*, *Yeshilbashian Y.*, *Avetisyan K.* Subwords-Only Alternatives to fastText for Morphologically Rich Languages // Programming and Computer Software. — 2021. — Т. 47, № 1. — С. 56—66. — DOI: https://doi.org/10.1134/S0361768821010059.
6. *Ешилбашян Е.*, *Асатрян А.*, *Гукасян Ц.* Поиск заимствований в армянских текстах путем внутреннего стилометрического анализа // Труды Института системного программирования РАН. — 2021. — Т. 33, № 1. — С. 209—224. — DOI: https://doi.org/10.15514/ISPRAS-2021-33(1)-14.
7. *Avetisyan K.*, *Ghukasyan T.* Word Embeddings for the Armenian Language: Intrinsic and Extrinsic Evaluation. // “Vestnik” Scientific Journal of Russian-Armenian University. — 2019. — № 1. — С. 59—72.