ISSN: 2224-5227 (Print)
ISSN: 2518-1483 (Online)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№3
2025

ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

3 (355)

JULY – SEPTEMBER 2025

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).

Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.

Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*

Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*

Мерзімділігі: *жылына 4 рет.*

http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*
Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**
Тематическая направленность: *информационно-коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*
Периодичность: *4 раза в год.*
http://www.physico-mathematical.kz/index.php/en/

CONTENTS

МАЗМҰНЫ

СОДЕРЖАНИЕ

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 188–203

https://doi.org/10.32014/2025.2518-1726.372

МРНТИ 28.23.37
УДК 004.89

D. Oralbekova[1*], A. Akhmediyarova[2], D. Kassymova[3], Z. Alibiyeva[2], 2025.
[1]Institute of information and computational technologies, Almaty, Kazakhstan;
[2] Satbayev University, Almaty, Kazakhstan;
[3] ALT University, Almaty, Kazakhstan.
E-mail: dinaoral@mail.ru

RESEARCH ON LINGUISTIC ANALYSIS METHODS FOR IDENTIFYING AND EXTRACTING TEXT DATA IN THE KAZAKH LANGUAGE

Oralbekova Dina — PhD, senior researcher, Institute of information and computational technologies, Almaty, Kazakhstan,
E-mail: dinaoral@mail.ru, ORCID ID: https://orcid.org /0000-0003-4975-6493;
Akhmediyarova Ainur — PhD, professor, Satbayev University, Almaty, Kazakhstan,
E-mail: a.akhmediyarova@satbayev.university, ORCID ID: https://orcid.org/0000-0003-4439-7313;
Kassymova Dinara — PhD, assistant professor, ALT University, Almaty, Kazakhstan,
E-mail: d.kassymova@alt.edu.kz, ORCID ID: https://orcid.org/0000-0001-6152-8317;
Alibiyeva Zhibek — PhD, Associate professor, Satbayev University, Almaty, Kazakhstan,
E-mail: zh.alibiyeba@satbayev.university, ORCID ID: https://orcid.org/0000-0001-9565-5621.

Abstract. This paper examines modern linguistic analysis methods applied to the processing of the Kazakh language for the purpose of automatic identification and extraction of textual information. Special attention is given to morphological, syntactic, and semantic analysis and their adaptation to the specific features of the Kazakh language, which is classified as an agglutinative language and is characterized by flexible word order. These features create certain challenges when applying traditional approaches designed for languages with fixed word order, such as English. The study analyzes contemporary approaches, including finite-state machine methods, statistical models, deep neural networks, and transformer-based architectures. It reviews existing software tools such as HFST, Apertium, KazNERD, BeeBERT, and Kaz-RoBERTa, as well as other models specifically adapted for languages with complex morphological structures. Their potential and limitations are assessed in the context of Kazakh text processing. Particular focus is placed on the accuracy of morphological analysis, the models' robustness to polysemy, and their ability to handle rare and complex word forms. The paper also discusses practical applications of modern NLP solutions for the Kazakh

language — in machine translation systems, automatic text classification, named entity recognition, and sentiment analysis. Concrete examples of model usage in the educational and legal domains are presented. Finally, the paper provides recommendations for developing national text corpora, advancing morphological analysis tools, and further exploring the integration of different methodological approaches to improve the quality of Kazakh language processing in NLP tasks.

Keywords: Kazakh language, Transformer, morphological analysis, syntactic and semantic analysis, NLP, pretrained models

Д. Оралбекова[1*], А. Ахмедиярова[2], Д. Қасымова[3], Ж. Алибиева[2], 2025.
[1] Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан;
[2] Satbayev университеті, Алматы, Қазақстан;
[3] М. Тынышпаев атындағы ALT университеті, Алматы, Қазақстан.
E-mail: dinaoral@mail.ru

ҚАЗАҚ ТІЛІНДЕГІ МӘТІНДІК АҚПАРАТТЫ АНЫҚТАУ ЖӘНЕ ОНЫ ШЫҒАРЫП АЛУ ҮШІН ЛИНГВИСТИКАЛЫҚ ТАЛДАУ ӘДІСТЕРІН ЗЕРТТЕУ

Оралбекова Дина — PhD, аға ғылыми қызметкер, Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,
E-mail: dinaoral@mail.ru, ORCID ID: https://orcid.org/0000-0003-4975-6493;
Ахмедиярова Айнұр — PhD, профессор, Satbayev университеті, Алматы, Қазақстан,
E-mail: a.akhmediyarova@satbayev.university, ORCID ID: https://orcid.org/0000-0003-4439-7313;
Қасымова Динара — PhD, ассистент-профессор, М. Тынышпаев атындағы ALT университеті. Алматы, Қазақстан,
E-mail: d.kassymova@alt.edu.kz, ORCID ID: https://orcid.org/0000-0001-6152-8317;
Алибиева Жибек — PhD, қауымдастырылған профессор, Satbayev университеті, Алматы, Қазақстан,
E-mail: zh.alibiyeva@satbayev.university, ORCID ID: https://orcid.org/0000-0001-9565-5621.

Аннотация. Бұл мақалада қазақ тілін өңдеуге бағытталған заманауи лингвистикалық талдау әдістері қарастырылды. Мәтіндік ақпаратты автоматты түрде анықтау және шығарып алу мақсатында қолданылатын тәсілдерге ерекше назар аударылды. Морфологиялық, синтаксистік және семантикалық талдау түрлері мен олардың қазақ тіліне бейімделуі егжей-тегжейлі сипатталды. Қазақ тілі агглютинативті тілдер қатарына жатқандықтан сөз тәртібінің еркіндігімен ерекшеленеді. Мұндай ерекшеліктер ағылшын тілі сияқты сөз тәртібі қатаң тілдерге арналған дәстүрлі тәсілдерді қолдануда белгілі бір қиындықтар туғызады. Зерттеуде қазіргі таңдағы тәсілдер қарастырылған, оның ішінде атап айтқанда келесілер келтірілген: ақырлы автоматтар әдістері, статистикалық модельдер, терең нейрондық желілер және трансформер негізіндегі архитектуралар талданады. HFST, Apertium, KazNERD, BeeBERT және Kaz-RoBERTa сияқты бағдарламалық құралдармен қатар, күрделі

морфологиялық құрылымдарға бейімделген басқа да модельдерге мен тәсілдерге шолу жасалды. Бұл құралдардың қазақ мәтінін өңдеу контекстіндегі мүмкіндіктері, артықшылықтары мен шектеулері сарапталды. Морфологиялық талдаудың дәлдігіне, модельдердің көпмағыналылыққа төзімділігіне және сирек әрі күрделі сөз тұлғаларын өңдеу қабілетіне ерекше назар аударылды. Сонымен қатар, қазіргі NLP шешімдерінің қазақ тіліне арналған практикалық қолдану салалары машиналық аударма жүйелері, мәтіндерді автоматты түрде жіктеу, атаулы мәндерді тану және тональдікті талдау мәселелері қозғалады. Модельдердің білім беру және құқық салаларында қолданылуына нақты мысалдары келтіріледі. Мақала соңында ұлттық мәтін корпустарын құру және өңдеу, морфологиялық талдау құралдарын жетілдіру, сондай-ақ қазақ тілін өңдеудің сапасын арттыру мақсатында түрлі әдістемелік тәсілдерді біріктіру бойынша ұсыныстар берілген.

Түйін сөздер: қазақ тілі, Transformer, морфологиялық талдау, синтаксистік және семантикалық талдау, NLP, алдын ала үйретілген модельдер

Д. Оралбекова[1]*, А. Ахмедиярова[2], Д. Касымова[3], Ж. Алибиева[2], 2025.
[1] Институт информационных и вычислительных технологий,
Алматы, Казахстан;
[2] Satbayev University, Алматы, Казахстан;
[3] ALT университет имени М. Тынышпаева, Алматы, Казахстан.
E-mail: dinaoral@mail.ru

ИССЛЕДОВАНИЕ МЕТОДОВ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ДЛЯ ВЫЯВЛЕНИЯ И ИЗВЛЕЧЕНИЯ ТЕКСТОВЫХ ДАННЫХ НА КАЗАХСКОМ ЯЗЫКЕ

Оралбекова Дина — PhD, старший научный сотрудник, Институт информационных и вычислительных технологий, Алматы, Казахстан,
E-mail: dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493;
Ахмедиярова Айнур — PhD, профессор, Satbayev Университет, Алматы, Казахстан,
E-mail: a.akhmediyarova@satbayev.university, https://orcid.org/0000-0003-4439-7313;
Касымова Динара — PhD, ассистент-профессор, ALT университет имени М. Тынышпаева, Алматы, Казахстан,
E-mail: d.kassymova@alt.edu.kz, https://orcid.org/0000-0001-6152-8317;
Алибиева Жибек — PhD, ассоциированный профессор, Satbayev Университет, Алматы, Казахстан,
E-mail: zh.alibiyeva@satbayev.university, https://orcid.org/0000-0001-9565-5621.

Аннотация. В данной статье рассматриваются современные методы лингвистического анализа, применяемые для обработки казахского языка, с целью автоматического выявления и извлечения текстовой информации. Особое внимание уделяется морфологическому, синтаксическому и семантическому анализу, а также их адаптации к особенностям казахского языка, который относится к агглютинативным языкам и характеризуется свободным порядком

слов. Это создаёт определённые трудности при применении традиционных подходов, разработанных для языков с фиксированным порядком слов, таких как английский. В исследовательской работе анализируются современные подходы, включая методы на основе конечных автоматов, статистические модели, глубокие нейронные сети и трансформерные архитектуры. Рассматриваются существующие программные инструменты, такие как HFST, Apertium, KazNERD, BeeBERT и Kaz-RoBERTa и другие модели, специально адаптированные для языков со сложной морфологической структурой, а также их потенциал и ограничения в контексте обработки казахских текстов. Особое внимание уделяется вопросам точности морфологического анализа, устойчивости моделей к полисемии, а также способности справляться с редкими и сложными словоформами. Также обсуждаются практические области применения современных NLP-решений для казахского языка — в системах машинного перевода, автоматической классификации текстов, извлечении именованных сущностей и анализе тональности. Представлены конкретные примеры применения моделей в образовательной и юридической сферах. В заключении даны рекомендации по созданию национальных текстовых корпусов, развитию инструментов морфологического анализа, а также дальнейшему исследованию интеграции различных методологических подходов для повышения качества обработки казахского языка в задачах NLP.

Ключевые слова: казахский язык, Transformer, морфологический анализ, синтаксический и семантический анализ, NLP, предобученные модели

Благодарности. *Данное исследование финансировалось Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант BR24993166).*

Введение. Обработка естественного языка (NLP) для казахского языка представляет собой ряд уникальных задач, обусловленных его агглютинативной природой и свободным порядком слов. Эти особенности требуют разработки специализированных методов лингвистического анализа для выявления и извлечения текстовых данных. Современные подходы, включая методы глубокого обучения, трансформерные модели и статистические техники, открывают возможности для создания высокоточных инструментов обработки текста. Однако ограниченная доступность размеченных данных и сложность грамматической структуры языка затрудняют реализацию подобных решений.

В данном исследовании основное внимание уделяется анализу современных методов лингвистического анализа — морфологического, синтаксического и семантического — и их адаптации к особенностям казахского языка.

NLP — это ключевое направление в области искусственного интеллекта и лингвистики, предоставляющее передовые методы автоматического анализа текста для таких задач, как извлечение информации, машинный перевод и анализ тональности. Однако для казахского языка, как и для многих других

языков с ограниченными ресурсами, разработка эффективных NLP-решений сопряжена со значительными трудностями.

Казахский язык относится к классу агглютинативных языков, в которых грамматическое значение выражается посредством аффиксов, присоединяемых к корню слова. Это требует создания специализированных морфологических анализаторов, способных точно интерпретировать сложные и многокомпонентные словоформы. Кроме того, свободный порядок слов в предложениях создает дополнительные сложности для синтаксического и семантического анализа, поскольку традиционные методы, разработанные для языков с фиксированным порядком слов, часто оказываются неэффективными.

Современные технологии машинного обучения, такие как глубокие нейронные сети и трансформерные модели, открывают новые возможности для обработки казахских текстов. Модели, такие как BeeBERT и KazNERD (Yeshpanov et al., 2022), демонстрируют заметный прогресс в анализе текстов, однако их производительность по-прежнему зависит от наличия крупных размеченных корпусов, которые в случае казахского языка пока остаются ограниченными.

Цель данного исследования — провести обзор и систематизацию существующих методов лингвистического анализа, адаптировать их под казахский язык и оценить применимость различных подходов, включая конечные автоматы, статистические модели и трансформеры. Это позволит определить современные достижения в обработке казахского языка и обозначить перспективные направления для дальнейшего развития, такие как создание новых текстовых корпусов и интеграция различных методологических подходов.

Материалы и методы исследования. Морфологический анализ — это процесс разбиения слова на его составные части (корень и аффиксы) и определение их грамматических значений. Для агглютинативных языков, таких как казахский, морфологический анализ особенно важен, поскольку грамматические значения передаются с помощью многочисленных аффиксов, присоединяемых к корню слова.

Методы, основанные на правилах и словарях

Метод, основанный на правилах и словарях, представляет собой базовый подход к вычислительному морфологическому анализу. В словарной базе хранятся базовые формы слов (леммы), а набор правил описывает порядок присоединения аффиксов и их взаимодействие (Jurafsky et al., 2019).

Словари, являясь центральным элементом данного подхода, содержат морфологические характеристики лексем, такие как часть речи и основные грамматические признаки (род, число, падеж и т. д.) (Haspelmath et al., 2013). При проверке слова по словарю анализатор применяет правила для определения структуры и значения слова (Kaplan et al., 1994).

К ключевым инструментам данной категории относятся TRMorph и Apertium. TRMorph — морфологический анализатор для турецкого языка,

обеспечивающий точный анализ за счёт строгих правил и структурированных словарей (Kim, 2024). Аналогично, Apertium — это программное обеспечение с открытым исходным кодом, предназначенное для морфологического анализа и машинного перевода, поддерживающее агглютинативные языки, включая казахский (Forcada et al., 2011).

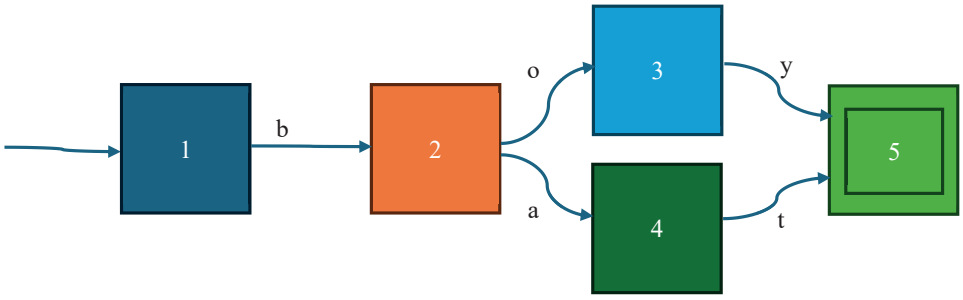
Подход, основанный на правилах и словарях, обладает рядом преимуществ, таких как высокая точность для слов, присутствующих в словаре, и относительная простота реализации для языков с хорошо изученной морфологией. Однако он также имеет ограничения — в частности, слабая способность обрабатывать неизвестные слова и значительные затраты на разработку словарей и правил.

Конечные автоматы

Метод конечных автоматов применяется для моделирования регулярных структур и морфологических процессов, что делает его особенно полезным для анализа агглютинативных языков. Конечные автоматы могут быть детерминированными (DFA) или недетерминированными (NFA), что позволяет представлять сложные грамматические системы с помощью формальных языков и регулярных выражений (Boyd et al., 2021).

Конечный автомат представляет собой графовую структуру, где узлы соответствуют состояниям, а переходы между ними моделируют морфологические преобразования. Морфологический анализатор обрабатывает входное слово, проходя по состояниям автомата и применяя заранее заданные правила для извлечения морфологических характеристик слова (Beesley et al., 2003) (рис. 1).

Рисунок 1. Общая структура конечного автомата



Наиболее широко используемыми инструментами в этой области являются HFST и FOMA.

HFST (Helsinki Finite-State Transducer) — это платформа для построения конечных автоматов, широко применяемая для анализа агглютинативных языков, включая казахский. HFST поддерживает интеграцию с другими инструментами, такими как Apertium, и обеспечивает высокую производительность благодаря оптимизированным алгоритмам (Lindén et al., 2011).

FOMA — гибкий и легковесный инструмент для разработки и тестирования конечных автоматов. Он используется для создания сложных морфологических моделей и поддерживает различные форматы ввода/вывода, что делает его универсальным решением (Hulden, 2009).

Метод конечных автоматов обладает высокой эффективностью при обработке регулярных структур, таких как порядок аффиксов в агглютинативных языках. Его адаптивность делает его подходящим для построения компактных и быстрых морфологических анализаторов. Однако у него имеются и недостатки: ограниченная гибкость при обработке нерегулярных форм и исключений, а также необходимость глубоких знаний в области формальных языков, морфологии агглютинативных языков и программирования, что может создавать сложности для разработчиков (Manohar et al., 2022).

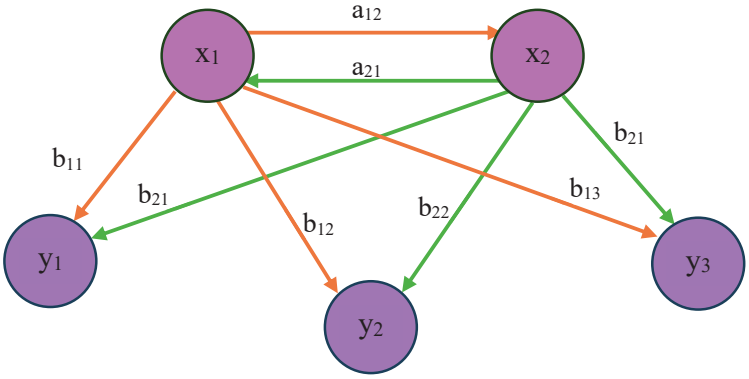
Статистические модели

Методы статистического анализа основаны на вероятностных подходах для предсказания морфологических тегов, что делает их адаптивными и пригодными для обработки языков с большими корпусами данных. В отличие от систем, основанных на правилах, эти модели не ограничены заранее заданными правилами, что позволяет им обрабатывать неизвестные слова и учиться на реальных текстах (Koosha et al., 2022).

Основной принцип статистических моделей заключается в использовании распределений вероятностей для анализа последовательностей слов и их морфологических тегов. Такие методы, как скрытые марковские модели (HMM) и условные случайные поля (CRF), используют вероятностное графовое моделирование для захвата сложных зависимостей между аффиксами и лексемами (Fraser, 2008).

HMM оценивают наиболее вероятную последовательность тегов для заданного входа на основе вероятностей переходов. Однако эффективность этих моделей ограничивается предположением, что каждое состояние зависит только от предыдущего (Wang, 2022) (рис. 2).

Рисунок 2. Пример скрытой марковской модели



В отличие от HMM, CRF предлагают более продвинутый подход, учитывая глобальные зависимости и взаимодействия между элементами входной последовательности. Это делает их более эффективными для обработки сложных языков, таких как казахский (Oralbekova et al., 2022).

Методы статистического анализа имеют ряд преимуществ: они способны обрабатывать неизвестные слова и использовать вероятностные рассуждения, что делает их весьма адаптивными к новым данным. Для языков с большими размеченными корпусами, таких как английский или китайский, они обеспечивают высокую точность морфологического анализа.

Основными недостатками статистических моделей являются их зависимость от большого объема размеченных данных, а также необходимость значительных вычислительных ресурсов и знаний в области машинного обучения для обучения и тонкой настройки моделей.

Методы глубокого обучения

Современные методы глубокого обучения используют нейронные сети для автоматического извлечения морфологических закономерностей из данных. Такой подход продемонстрировал высокую эффективность благодаря способности нейросетей обрабатывать сложные языковые зависимости и адаптироваться к особенностям агглютинативных языков, таких как казахский.

Глубокие нейронные сети, такие как LSTM (долгая краткосрочная память) и GRU (затворные рекуррентные единицы), широко применяются для обработки последовательных данных. Эти модели особенно хорошо подходят для задач, требующих учета порядка элементов, таких как анализ последовательностей аффиксов в словах. LSTM и GRU способны удерживать долгосрочные зависимости и учитывать контекст при анализе морфологической структуры (Vennerød et al., 2021).

Более продвинутые архитектуры, такие как трансформеры (включая BERT и Kazakh-BERT), используют механизмы внимания для параллельной обработки всего контекста слова. Это позволяет эффективно анализировать сложные морфологические структуры и контекстуальные вариации даже в языках с богатой морфологией (Vaswani et al., 2017).

Методы глубокого обучения обладают рядом преимуществ: они достигают высокой точности при наличии больших объемов данных, могут моделировать сложные морфологические закономерности, включая нерегулярные формы и редкие комбинации аффиксов. Кроме того, они способны обрабатывать слова с несколькими возможными интерпретациями (Devlin et al., 2019).

Основные недостатки этих методов — высокая вычислительная стоимость и зависимость от больших размеченных корпусов. Обучение таких моделей требует значительных аппаратных ресурсов, а также тщательной подготовки текстовых данных.

Методы синтаксического анализа

Синтаксический анализ (парсинг) играет ключевую роль в обработке текстов, особенно для языков со свободным порядком слов, таких как

казахский. Он позволяет выявить структурные зависимости между словами в предложении и сформировать их грамматическое представление. Методы синтаксического анализа можно условно разделить на два подхода: аналитический (на основе правил) и статистический (на основе моделей машинного обучения).

Зависимостный парсинг (Dependency Parsing)

Зависимостный парсинг представляет собой метод, при котором строится графовая структура, в которой узлы соответствуют словам, а ребра обозначают синтаксические связи между ними. Такая структура наглядно показывает, какие элементы предложения зависят друг от друга и как они связаны.

Основная цель метода — определить зависимости между словами в предложении. Например, подлежащее и сказуемое связаны отношением зависимости. Этот подход особенно эффективен для языков со свободным порядком слов, таких как казахский, где грамматические функции передаются морфологически, а не фиксированной позицией в предложении.

Существуют два основных алгоритма зависимостного анализа: 1) Переходный анализ (Transition-Based Parsing) — строит дерево зависимостей пошагово. Каждый шаг определяет действие (например, добавить ребро или перейти к следующему слову), что делает алгоритм быстрым и эффективным. Однако он чувствителен к ошибкам на ранних этапах. 2) Графовый анализ (Graph-Based Parsing) — строит глобально оптимальное дерево зависимостей, формулируя задачу как нахождение максимального остовного дерева в графе. Обеспечивает высокую точность за счёт рассмотрения всех возможных связей, но требует больших вычислительных ресурсов.

Фразовая структура (Constituency Parsing)

Фразовый анализ представляет предложение как иерархическую структуру, где каждый узел соответствует грамматической единице: фразе, придаточному предложению или всему предложению. Метод исходит из предположения, что каждое предложение можно разделить на более мелкие грамматические компоненты.

Дерево фразовой структуры показывает, как слова объединяются в более крупные синтаксические единицы. Например, глагольная фраза (VP) может включать глагол и его дополнения. Этот метод полезен для глубокого синтаксического анализа и широко применяется в машинном переводе и лингвистическом анализе.

Алгоритмы фразового анализа:

Алгоритм CKY (Cocke–Kasami–Younger) — используется для построения деревьев разбора на основе контекстно-свободных грамматик (CFG). Применяет динамическое программирование, обеспечивая эффективность для языков с фиксированным порядком слов.

Глубокие нейросети — современные подходы используют LSTM и трансформеры для предсказания фразовых структур. Эти модели способны

учитывать контекст всего предложения, что особенно полезно для языков со свободным порядком слов, таких как казахский (Oralbekova et al., 2024).

Статистический и нейросетевой парсинг

Современные методы синтаксического анализа всё чаще используют глубокие нейронные сети для автоматического выявления синтаксических паттернов из больших текстовых корпусов. Эти методы достигают высокой точности и адаптивности, что делает их пригодными для анализа текстов на различных языках, включая казахский.

Ключевыми инструментами являются BiLSTM и трансформеры. BiLSTM (двунаправленная LSTM) эффективно обрабатывает последовательные зависимости, анализируя как предшествующий, так и следующий контекст. Трансформеры (включая BERT и GPT) моделируют сложные синтаксические структуры с помощью механизмов внимания, позволяя учитывать весь контекст предложения (Narejo et al., 2024).

Методы семантического анализа

Семантический анализ включает извлечение смысла из текста, включая определение значений слов, выявление связей между ними и интерпретацию контекста. Он играет ключевую роль в таких задачах, как машинный перевод, извлечение информации и анализ тональности.

Векторные представления слов (Word Embeddings)

Методы векторного представления слов преобразуют слова в числовые векторы, отражающие их значение в контексте. Эти методы основаны на дистрибутивной гипотезе, согласно которой слова с близкими значениями появляются в сходных контекстах. Векторные представления обучаются на основе анализа совместной встречаемости слов в текстах, что делает их универсальным инструментом обработки текстов.

Ключевые методы:

Word2Vec — создаёт векторы слов с использованием двух архитектур:

- CBOW (непрерывный мешок слов) предсказывает текущее слово по его окружению.

- Skip-Gram предсказывает окружающие слова по текущему.

Word2Vec хорошо подходит для кластеризации слов и анализа семантического сходства. Например, слова «қала» (город) и «ауыл» (село) будут иметь близкие векторы.

GloVe (Global Vectors for Word Representation) учитывает глобальную статистику совместной встречаемости слов в больших корпусах. Обучается на матрице встречаемости и хорошо справляется с задачами семантического сравнения.

FastText в отличие от Word2Vec и GloVe, работает на уровне подслов, анализируя последовательности символов в словах. Это особенно эффективно для морфологически богатых языков, таких как казахский, где аффиксы существенно меняют значение слова. Например, «мектеп» (школа) и «мектептер» (школы) будут иметь схожие представления.

Основные преимущества этих моделей — простота реализации, быстрая скорость обучения и интуитивная интерпретация: семантически близкие слова отображаются близко в векторном пространстве. Однако они имеют ограничения: 1) Каждому слову присваивается один вектор, что затрудняет обработку многозначных слов. 2) Сложно учитывать сложные контекстные зависимости, особенно в длинных предложениях.

Семантическое ролевое аннотирование (Semantic Role Labeling, SRL)

Семантическое ролевое аннотирование — это процесс определения ролевой структуры предложения, при котором каждому слову или фразе присваивается определённая функция в контексте действия. Этот метод позволяет распознавать субъекты, объекты, действия и другие компоненты, формируя насыщенное семантическое представление текста.

SRL направлен на определение семантических ролей в предложении. Например, в предложении «Али прочитал книгу» SRL аннотирует «Али» как субъект, «прочитал» — как действие, а «книгу» — как объект. Таким образом, SRL является мощным инструментом для анализа сложных синтаксических структур, особенно в задачах извлечения информации.

Основные подходы к SRL

Модели на основе правил. Этот подход опирается на заранее определённые шаблоны и грамматические правила для присвоения семантических ролей. Он прост в реализации, но обладает низкой гибкостью и требует значительной ручной настройки под каждый язык.

Глубокое обучение. Современные методы используют RNN, LSTM и трансформеры для автоматического аннотирования ролей. Эти модели способны улавливать как локальные, так и глобальные контексты, обеспечивая высокую точность даже в сложных структурах предложений (Mamyrbayev, 2023).

SRL особенно полезен для извлечения информации в таких областях, как анализ требований, обработка юридических документов и системы вопросов-ответов. Он хорошо подходит для длинных предложений, где необходимо выявить сложные отношения между словами. Однако у метода есть ограничения: он требует больших размеченных корпусов, что затрудняет его применение в условиях ограниченных языковых ресурсов. Кроме того, модели глубокого обучения могут испытывать трудности при обработке очень длинных текстов, что увеличивает вычислительные затраты и требует дополнительных ресурсов (Onan, 2023).

Пример применения: SRL может быть использован для извлечения требований из технической документации. В предложении «Система должна позволять пользователям загружать файлы» SRL аннотирует «система» как субъект, «должна позволять» — как действие, а «пользователей» и «файлы» — как объекты.

Результаты и обсуждение. Модели на основе трансформеров, такие как BERT, RoBERTa, T5, BART и GPT, стали прорывными инструментами

в области обработки естественного языка благодаря способности учитывать как левый, так и правый контекст слова. Эти модели обучаются на задачах маскированного языкового моделирования (MLM) и предсказания следующего предложения (NSP), что позволяет им эффективно анализировать как локальные, так и глобальные зависимости в тексте (Wang et al., 2024).

Модели глубокого обучения широко применяются в задачах извлечения смысловой информации. Они автоматизируют сложные процессы анализа текста, такие как перевод, классификация, распознавание именованных сущностей (NER) и анализ тональности, что делает их незаменимыми при обработке казахского языка (табл. 1).

Машинный перевод. Модели, такие как BERT, значительно улучшают качество перевода казахских текстов. Предобученные трансформеры могут адаптироваться к агглютинативной морфологии языка и свободному порядку слов. Например, многоязычный BERT в сочетании с механизмами внимания улавливает контекст и грамматические структуры, повышая точность и естественность перевода.

Классификация текста. FastText, способный обрабатывать текст на уровне символов, хорошо подходит для классификации казахских документов. Модель эффективно справляется с морфологическими особенностями языка, обрабатывая тексты на различные темы. Например, FastText успешно применяется для автоматической категоризации документов по таким направлениям, как образование, политика и наука.

Распознавание именованных сущностей (NER). KazNERD — модель, специально разработанная для казахского языка, эффективно распознаёт сущности, такие как имена, организации и географические названия. Интеграция глубокого обучения с трансформерами обеспечивает точное извлечение сложных лингвистических конструкций, учитывая морфологическое богатство языка. Например, KazNERD используется при анализе юридических текстов, где важно идентифицировать участников и наименования организаций.

Семантическое сходство. Методы векторного представления текста, такие как Word2Vec, широко применяются для оценки семантического сходства. Эти подходы измеряют близость значений между двумя текстами или словами. Например, Word2Vec используется для кластеризации казахских текстов, группируя схожие документы на основе их содержания.

Анализ тональности. Модель BERT показывает высокую точность при решении задач анализа тональности на казахском языке. Она анализирует контекст слов в предложении, позволяя точно определять эмоциональную окраску (положительную, отрицательную или нейтральную). Такой подход используется, например, при анализе отзывов пользователей на казахском языке, что помогает компаниям понимать обратную связь и улучшать клиентский сервис.

Таблица №1 – Обзор инструментов лингвистического анализа для казахского языка

	Инструмент / Модель	Описание	Преимущества	Недостатки
Морфологический анализ	Apertium	Программное обеспечение с открытым исходным кодом для машинного перевода и морфологического анализа. Поддерживает казахский язык и использует словари на основе правил.	Подходит для базового анализа текста и перевода.	Ограничен фиксированными правилами, низкая эффективность при работе со сложными словоформами.
	HFST	Helsinki Finite-State Transducer, предназначенный для агглютинативных языков.	Эффективные модели на основе конечных автоматов, поддержка универсальных морфологических описаний.	Требует значительных усилий по настройке правил.
	KazNERD	Система распознавания именованных сущностей для казахского языка, интегрирует морфологический анализ.	Учитывает морфологические особенности языка, высокая точность.	Ограничена задачами распознавания именованных сущностей.
	BeeBERT	Адаптированная версия BERT для казахского языка.	Улавливает морфологические и синтаксические особенности, сохраняет высокую точность даже при малом объеме данных.	Требует высоких вычислительных ресурсов.
Синтаксический анализ	Stanford Parser	Использует универсальные грамматики зависимостей для анализа текста.	Поддерживает множество языков, включая казахский.	Сложно настраивается для агглютинативных языков.
	MSTParser	Реализует графовый подход к синтаксическому разбору зависимостей.	Точный и гибкий при анализе сложных синтаксических структур.	Медленная обработка больших корпусов.
	Berkeley Parser	Поддерживает несколько языков, использует статистические модели для парсинга по составляющим.	Высокая точность для агглютинативных языков.	Ограниченное количество моделей для языков с низкими ресурсами.
	Stanford Constituency Parser	Применяет статистические методы для синтаксического анализа.	Подходит для языков с фиксированным порядком слов.	Менее эффективен для казахского языка из-за его свободного порядка слов.

	UDPipe	Инструмент на основе BiLSTM, поддерживает токенизацию, морфологический и синтаксический анализ.	Универсальное решение с поддержкой многих языков.	Зависит от качества обучающих данных.
	SpaCy	Современная библиотека NLP для обработки текста.	Быстрая, легко интегрируется.	Ограниченная поддержка казахского языка.
Семантический анализ	KazSemEval	Платформа для оценки семантических задач на казахском языке, включая извлечение связей и анализ значений слов.	Специально разработана для казахского языка, учитывает лингвистические и культурные особенности.	Недостаточные ресурсы для решения сложных семантических задач.
	Kaz-RoBERTa	Модифицированная версия RoBERTa, адаптированная под казахский язык. Поддерживает широкий спектр задач NLP.	Высокая точность в задачах классификации текста, анализа тональности и извлечения информации.	Требует больших обучающих корпусов и значительных вычислительных ресурсов.

Нерешённые проблемы и направления дальнейших исследований

Одной из ключевых проблем в обработке казахского языка является ограниченность размеченных данных. Современные методы глубокого обучения, такие как трансформеры, требуют масштабных корпусов для обучения, что затрудняет разработку эффективных моделей. Будущие исследования должны быть сосредоточены на создании и аннотировании крупных текстовых корпусов, включая специализированные области, такие как право и медицина.

Агглютинативная структура казахского языка, при которой к корню слова присоединяется множество аффиксов, усложняет интеграцию морфологического и синтаксического анализа. Аналитические инструменты должны одновременно учитывать морфологические преобразования и синтаксические зависимости, что требует высоких вычислительных ресурсов и сложной настройки моделей.

Для языков со свободным порядком слов, таких как казахский, традиционные методы синтаксического анализа сталкиваются с трудностями. Грамматические функции определяются морфологическими маркерами, а не позицией слова в предложении, что требует разработки языкоориентированных моделей, способных учитывать эти особенности.

Несмотря на то, что современные модели, такие как BERT, эффективно улавливают контекст, они всё ещё сталкиваются с проблемой полисемии (множественности значений слов). Эта проблема особенно актуальна для казахского языка, где значение слова зависит от морфологии и контекста. В

будущем исследования должны быть направлены на улучшение разрешения полисемии, например, путём интеграции мультимодальных подходов.

Высокие вычислительные требования моделей глубокого обучения ограничивают их практическое применение. В дальнейшем необходимо разрабатывать оптимизированные и легковесные модели для внедрения в условия с ограниченными ресурсами. Передобучение на больших многоязычных корпусах, таких как Multilingual-BERT, с последующей донастройкой на меньших казахских выборках, может помочь решить проблему нехватки данных. Этот подход уже продемонстрировал свою эффективность в задачах синтаксического анализа.

Заключение. В данной работе представлен обзор современных методов лингвистического анализа для обработки текстов на казахском языке. Исследование охватывает подходы к морфологическому, синтаксическому и семантическому анализу, включая методы на основе глубоких нейронных сетей и трансформеров.

В анализе подчеркнуты значительные достижения в данной области, включая разработку специализированных инструментов и моделей, а также выявлены основные нерешённые проблемы. Ключевыми из них являются: нехватка размеченных данных, сложность обработки агглютинативной морфологии и высокие вычислительные затраты моделей глубокого обучения. Для преодоления этих барьеров необходима дальнейшая работа по развитию корпусов, адаптации моделей и созданию новых методов лингвистического анализа. Несмотря на существующие ограничения, результаты исследований показывают, что современные технологии обработки естественного языка могут значительно повысить эффективность автоматизированной обработки казахских текстов, открывая новые возможности для применения в образовании, праве и сфере искусственного интеллекта.

References

Boyd R.L., Schwartz H.A. (2021) Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. Journal of Language and Social Psychology, 40(1). — P. 21-41. https://doi.org/10.1177/0261927X20967028 (in English)

Beesley K.R., Karttunen L. (2003) Finite State Morphology, CSLI Studies in Computational Linguistics. Finite State Morphology, CSLI Studies in Computational Linguistics (in English)

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — P. 4171–4186 (in English)

Forcada M.L., et al. (2011) Apertium: A Free/Open-Source Platform for Rule-Based Machine Translation. Machine Translation, vol. 25, no. 2. — P. 127–144 (in English)

Fraser A.M. (2008) Hidden Markov Models and Dynamical Systems. Society for Industrial and Applied Mathematics, USA. — P. 144, ISBN 0898717744, 9780898717747 (in English)

Haspelmath M., Sims A. (2013) Understanding Morphology (2nd ed.). Routledge, 384 p., eBook ISBN 9780203776506. https://doi.org/10.4324/9780203776506 (in English)

Hulden M. (2009) Foma: a finite-state compiler and library. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session (EACL '09). Association for Computational Linguistics, USA, 2009. — P. 29–32 (in English)

Jurafsky D., Martin J.H. (2019) Logistic Regression. In: Speech and Language Processing, 3rd

Edition (Draft). — P. 75-93. https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf (in English)

Kaplan R.M., Kay M. (1994) Regular Models of Phonological Rule Systems. Computational Linguistics, vol. 20, no. 3. — P. 331–378 (in English)

Kim Y. (2024) On morphological requirements for auxiliary verb periphrasis in Turkish. Glossa: a journal of general linguistics. Vol. 9(1), doi: https://doi.org/10.16995/glossa.9771 (in English)

Koosha S., Mahyar A., Yaser A., Godarzi A., Javad. (2022) Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model (October 2022). International Journal of Science and Information System Research, Volume 12, Issue 9. — P. 20 - 44, 2022, Available at SSRN: https://ssrn.com/abstract=4251017 (in English)

Lindén K., Axelson E., Hardwick S., Pirinen T.A., Silfverberg M. (2011) HFST—Framework for Compiling and Applying Morphologies. In: Mahlow, C., Piotrowski, M. (eds) Systems and Frameworks for Computational Morphology. SFCM. Communications in Computer and Information Science, vol 100. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23138-4_5 (in English)

Mamyrbayev O., Wojcik W., Titova N., Pavlov S., Oralbekova D., Aitkazina A., Zhumazhan N. (2023) Development of a thermodynamic model for optimization of processes in crop production. Eastern-European Journal of Enterprise Technologies, 6(8 (126), — P. 25–34, 2023. https://doi.org/10.15587/1729-4061.2023.290294 (in English)

Manohar K., Jayan A.R., Rajan R. (2022) Mlphon: A Multifunctional Grapheme-Phoneme Conversion Tool Using Finite State Transducers. IEEE Access, vol. 10. — P. 97555-97575, doi: 10.1109/ACCESS.2022.3204403 (in English)

Onan A. (2023) SRL-ACO: A text augmentation framework based on semantic role labeling and ant colony optimization. J. King Saud Univ. Comput. Inf. Sci., vol. 35, 101611 (in English)

Oralbekova D., Mamyrbayev O., Othman M., Alimhan K., Zhumazhanov B., Nuranbayeva B. (2022) Development of CRF and CTC Based End-To-End Kazakh Speech Recognition System, in Nguyen, N.T., Tran, T.K., Tukayev, U., Hong, T.P., Trawiński, B., Szczerbicki, E. (eds) Intelligent Information and Database Systems. ACIIDS 2022. Lecture Notes in Computer Science, vol. 13757, Springer, Cham. [Online]. Available: https://doi.org/10.1007/978-3-031-21743-2_41 (in English)

Oralbekova D., Mamyrbayev O., Zhumagulova S., Zhumazhan N. (2024) A Comparative Analysis of LSTM and BERT Models for Named Entity Recognition in Kazakh Language: A Multi-classification Approach. In: Agarwal, N., Sakalauskas, L., Tukeyev, U. (eds) Modeling and Simulation of Social-Behavioral Phenomena in Creative Societies. Communications in Computer and Information Science, vol 2211. Springer, Cham. https://doi.org/10.1007/978-3-031-72260-8_10 (in English)

Rani Narejo K., Zan H., Oralbekova D., Parkash Dharmani K., Orken M., Mukhsina K. (2024) Enhancing Emoji-Based Sentiment Classification in Urdu Tweets: Fusion Strategies With Multilingual BERT and Emoji Embeddings. IEEE Access, vol. 12, pp. 126587-126600, doi: 10.1109/ACCESS.2024.3446897 (in English)

Vennerød C.B., Kjærran A., Bugge E.S. (2021) Long Short-term Memory RNN. ArXiv, abs/2105.06756 (in English)

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., A. Gomez N., Kaiser Ł., Polosukhin I. (2017) Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 2017. — P. 6000–6010 (in English)

Wang Y. (2022) Using Machine Learning and Natural Language Processing to Analyze Library Chat Reference Transcripts. Information Technology and Libraries. Vol. 41. 10.6017/ital.v41i3.14967 (in English)

Wang J., Huang J.X., Tu X., Wang J., Huang A.J., Laskar Md T.R., Bhuiyan A. (2024) Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. ACM Comput. Surv. Vol. 56, 7. — P. 33 https://doi.org/10.1145/3648471 (in English)

Yeshpanov R., Khassanov Y., Varol H. A. (2022) KazNERD: Kazakh Named Entity Recognition Dataset. In Proceedings of the Thirteenth Language Resources and Evaluation Conference. — P. 417–426 (in English)

Publication Ethics and Publication Malpractice
in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The Central Asian Academic Research Center LLP follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Ж.Ш.Әден*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 25.09.2025.
Формат 60x881/8. Бумага офсетная.
Печать –ризограф. 20,0 п.л. Заказ 3.

Национальная академия наук РК
050000, Алматы, ул. Кунаева 132