

УДК 81'33:004.912

**ЛЕММАТИЗАЦИЯ КАК ИНСТРУМЕНТ ОБРАБОТКИ
ЕСТЕСТВЕННОГО ЯЗЫКА: ТЕОРЕТИЧЕСКИЕ И ПРИКЛАДНЫЕ
АСПЕКТЫ**

Натаван Джамаладдин кызы Рахманова
*Институт языкознания им. И. Насими
Национальной Академии Наук Азербайджана
Баку, Азербайджан, natis-sdunina@mail.ru*

В статье рассматривается лемматизация как один из ключевых инструментов обработки естественного языка и морфологического анализа текста. Анализируются теоретические основы понятия леммы, особенности процесса лемматизации и её отличия от стемминга. Особое внимание уделяется применению лемматизации в агглютинативных языках, в частности в азербайджанский язык, а также её роли в автоматическом анализе текста, информационном поиске и электронной лексикографии. Делается вывод о том, что лемматизация является важной технологией, обеспечивающей более точную интерпретацию семантической структуры текста в современных системах обработки естественного языка.

Ключевые слова: лемматизация; обработка естественного языка; анализ текста; лемма; стемминг.

Термин «лемма» восходит к древнегреческому ‘λήμμα’, первоначально означавшему ‘предположение’ или ‘полученное’, что в контексте современной компьютерной лингвистики трансформировалось в понятие канонической

словарной формы слова [1, с. 35]. В лингвистике, если выразится более точно, «лемма – это каноническая форма лексемы, где лексема обозначает совокупность всех форм слова, имеющих одно и то же значение, а лемма – это конкретная форма, выбранная в качестве базовой для представления этой лексемы» [2]. Леммы имеют особое значение в языках с высокой степенью агглютинации, таких как азербайджанский язык.

«Лемматизация – это процесс определения леммы для данного слова, благодаря которому различные словоизменительные формы одного слова могут анализироваться как единый элемент. Этот процесс формирует набор лемм для лексической базы данных. Лемматизация рассматривается как переход от словоформ, найденных в корпусе текстов, к леммам, которые становятся заголовками словарных статей» [3]. Этот процесс с помощью морфологического анализа определяет контекст слова и его роль в предложении, приводя его к стандартной словарной форме. «Лемматизация используется в инструментах обработки естественного языка для различных языков на этапе предварительной обработки текстов, где различные формы слов приводятся к их нормальной форме. Кроме того, с точки зрения лингвистики она значительно облегчает автоматическое формирование семейств слов и особенно подготовку словарей» [4, с. 476].

Б.Т.Тахироглы в своем исследовании, отмечая преимущества лемматизации при обработке языкового корпуса турецкого языка, указывает, что «лемматизация экономит время и труд. Данный процесс позволяет исследователю- лингвисту автоматизированным образом обрабатывать огромное количество разных форм слов, в результате которого система автоматически приводит их к словарной форме». [4, с. 478].

Применяя лемматизацию тексты можно анализировать точнее. Потому, что при анализе частоты слов разные грамматические формы не считаются отдельно, в результате чего языковой материал анализируется очень легко, и можно выявить какие понятия и темы действительно важны в тексте или корпусе.

В отличие от метода стемминга (stemming), лемматизация не ограничивается простым удалением окончаний, а учитывает грамматические правила языка и позволяет получить морфологически корректную форму, существующую в словаре. Этот подход обеспечивает высокую точность анализа, особенно в языках со сложной структурой и грамматическими формами, изменяющимися в зависимости от контекста [5]. Например, в азербайджанском языке слово *kitablarımızdan* ‘из наших книг’ при стемминге может быть сокращено лишь до основы *kitab* ‘книга’, тогда как лемматизация анализирует морфологическую структуру слова, выделяя корень и грамматические аффиксы множественного числа, принадлежности и исходного падежа, и приводит его к словарной форме *kitab* ‘книга’. Аналогично, словоформа *gəlirdilər* ‘они приходили’ в результате лемматизации преобразуется в начальную форму глагола *gəlmək* ‘приходить’. Такой подход позволяет более точно анализировать тексты, объединяя различные грамматические формы одного слова в единую лексическую единицу.

С. Лукас и др. дают лемматизацию такое определение: «Стемминг является приближением к более общей задаче, называемой лемматизацией – определению базовой формы слова и объединению таких слов в одну группу. Однако, в отличие от простого отсечения окончания слова, лемматизация представляет собой более сложный алгоритм, который определяет происхождение слова и

возвращает только его лемму, то есть общую (словарную) форму слова. Лемматизация также способна определять контекст слова; например, существительное *saw* “пила” она оставит без изменений, а глагол *saw* “видел” преобразует в форму *see* “видеть”» [6, с. 258].

Благодаря такому подходу слова, представленные в различных грамматических формах, анализируются как единая единица, а системы поиска информации и автоматического анализа текста очищаются от лишних элементов. Т.Коркмаз и другие отмечают, что «процесс лемматизации, называемый в турецком языке анализом корневой формы, представляет собой приведение слова к его базовой единице. Он имеет сходные черты с процессом стемминга. В обоих случаях целью является получение корня слова. Однако лемматизация стремится выделить морфологический или семантический корень слова. Благодаря этому можно получить более точные результаты, устраняя лишние элементы, которые рассматриваются как ‘шум’» [7, с. 34]. В то же время данный методологический подход, за счёт приведения слов к их исходной форме в языках со сложной морфологической структурой, значительно повышает эффективность поиска информации и точность информационных баз. Этот подход, требующий более глубокого морфологического анализа по сравнению со стеммингом, определяет надёжную базовую форму слова, соответствующую грамматическим нормам языка, с учётом контекста и части речи слова.

Однако возможности лемматизации не ограничиваются исключительно задачами автоматической обработки текста. Как справедливо отмечает И. В. Балканов, «современные цифровые технологии уже сегодня демонстрируют эффективность интеллектуальных механизмов языковой коррекции и распознавания словоформ. Исследователь обращает внимание на то, что программные продукты, подобные Microsoft Word, способны автоматически обнаруживать орфографические ошибки, исправлять их или предлагать пользователю корректные варианты написания» [8, с. 3240]. В этой связи И.В. Балканов полагает, что аналогичный принцип может быть реализован и в электронной лексикографии: «словарь сможет автоматически выполнять операцию лемматизации, распознавая искомую словоформу даже при наличии ошибок в написании, и предлагать пользователю релевантные словарные статьи» [8, с. 3241].

Подобная позиция расширяет традиционное понимание лемматизации как исключительно технического инструмента морфологического анализа и позволяет рассматривать её как важный механизм оптимизации взаимодействия пользователя с электронными лексикографическими системами. Вместе с тем данный подход вызывает определённую научную дискуссию, поскольку чрезмерная автоматизация процессов интерпретации словоформ может привести к ошибочному определению леммы в условиях многозначности, омонимии или контекстуальной вариативности языка. Тем не менее очевидно, что интеграция алгоритмов лемматизации в электронные словари способствует повышению их адаптивности, функциональности и ориентированности на коммуникативные потребности пользователя.

Этот процесс основывается на словарной базе и сложных методах морфологического анализа, позволяющих системе определять часть речи слова. Так, в некоторых моделях для повышения эффективности лемматизации наряду с автоматической разметкой частей речи используются дополнительные алгоритмы

морфологического анализа, предназначенные для устранения грамматической неоднозначности. Особенно в агглютинативных языках, таких как Азербайджанский язык, применение лемматизации требует использования гибридных подходов (основанных на правилах и машинном обучении), учитывающих сложные взаимосвязи между корнем слова и аффиксами.

Однако следует отметить, что лемматизация, по сравнению со стеммингом, требует большего объёма вычислительных ресурсов и более сложных словарных баз, что делает её применение на больших массивах данных более затратным с точки зрения времени [9, с. 289]. Тем не менее высокая точность получаемых результатов и полное соответствие грамматической структуре языка превращают этот процесс в незаменимый инструмент при решении сложных задач обработки естественного языка, включая автоматический перевод и семантический анализ текста.

Вместе с тем в современной корпусной лингвистике всё более активно обсуждается вопрос о необходимости интеграции лемматизации в системы автоматического анализа текстовых корпусов и конкордансов. Так, К. Габилова, рассматривая методологические подходы корпусной лингвистики и новые области применения конкордансов, подчёркивает, что «цифровая обработка языковых данных требует не только формального сопоставления словоформ, но и их морфологической нормализации, обеспечивающей более точную репрезентацию семантических связей внутри корпуса» [10, с. 91]. Подобная позиция фактически расширяет функциональные границы лемматизации, выводя её за рамки исключительно технического этапа предобработки текста и превращая в один из ключевых инструментов интерпретации языковых данных.

В определённой степени данная точка зрения получает развитие и в исследовании К. Габиловой, посвящённом методике создания конкорданса для тюркских письменных памятников на материале эпоса Китаби-Деде Коркуд [11]. Исследователь отмечает, что «при работе с историческими и агглютинативными текстами формальное сопоставление словоформ без учёта их лемматизированной основы существенно осложняет выявление лексико-семантических связей и снижает точность корпусного анализа» [11]. Данное высказывание позволяет утверждать следующее: отказ от полноценной лемматизации в ряде случаев способен привести к фрагментарной интерпретации текста, тогда как использование морфологически ориентированных алгоритмов обеспечивает более целостное представление языковой структуры и семантики памятника.

Эти технологии играют решающую роль в автоматическом анализе текстов и оптимизации систем информационного поиска. Анализируемый эмпирический и теоретический материал позволяет нам сделать вывод о том, что лемматизация является не просто инструментом грамматического упрощения, но и фундаментальной технологией, позволяющей моделям обработки естественного языка более точно интерпретировать глубинный смысл текста и семантические связи между его элементами.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Коротаев, С. М. Historical semantics of the class concept in academic literature: the experience of quantitative analysis / С. М. Коротаев // Journal of Economic Sociology. – 2024. – Vol. 25, № 3. – P. 13-50. DOI: 10.17323/1726-3247-2024-3-13-50.

2. Sharipov, M. Development of a rule-based lemmatization algorithm through finite state machine for Uzbek language / M. Sharipov, O. Sobirov // arXiv. – 2022. DOI: 10.48550/arxiv.2210.16006.
3. Akhmetov, I. Highly language-independent word lemmatization using a machine-learning classifier / I. Akhmetov, A. Pak, I. Ualiyeva, A. Gelbukh // Computación y Sistemas. – 2020. – Vol. 24.
4. Tahiroğlu, B. T. Lemmatizasyon ve Türkçe için bir lematizasyon uygulaması: elemanTR / B. T. Tahiroğlu // RumeliDE Dil ve Edebiyat Araştırmaları Dergisi. – 2021. – № 24. – S. 475-486. DOI: 10.29000/rumelide.995343.
5. Габибова, К. А. Функциональность в стемминге: обработка сложных форм глаголов с учетом гармонии гласных (на основе азербайджанского языка) / К. А. Габибова // Сборник материалов круглого стола «Национальный корпус казахского языка – лингвистическая база для обучения большой языковой модели (LLM)» в рамках регулярной научно-языковой площадки «Аскар Жубановские чтения-4». – 2025. – С. 274-279. DOI: 10.5281/zenodo.18161701.
6. Lucas, C. Computer-assisted text analysis for comparative politics / C. Lucas, R. A. Nielsen, M. E. Roberts [et al.] // Political Analysis. – 2015. – Vol. 23, № 2. – P. 254-277. DOI: 10.1093/pan/mpu019.
7. Korkmaz, T. Analysis of whether news on the Internet is real or fake by using deep learning methods and the TF-IDF algorithm / T. Korkmaz, A. Çetinkaya, H. Aydın, M. A. Barışkan // International Advanced Researches and Engineering Journal. – 2021. – Vol. 5, № 1. – P. 31-41. DOI: 10.35860/iarej.779019.
8. Балканов, И. В. Лемматизация как средство оптимизации макроструктуры электронного словаря / И. В. Балканов // Филологические науки. Вопросы теории и практики. – 2023. – Т. 16, вып. 10. – С. 3237-3244.
9. Schofield, A. Comparing apples to apple: the effects of stemmers on topic models / A. Schofield, D. Mimno // Transactions of the Association for Computational Linguistics. – 2016. – Vol. 4. – P. 287-300. DOI: 10.1162/tacl_a_00099.
10. Habibova, K. Corpus linguistics and concordance: methodological approaches and new applications / K. Habibova // Linguistic Researches. – 2025. – № 1. – P. 89-95. DOI: [10.59849/2664-5432.2025.1.89](https://doi.org/10.59849/2664-5432.2025.1.89)
11. Габибова, К. Методика создания конкорданса для тюркских письменных памятников (на примере эпоса «Китаби-Деде Коркут») / К. Габибова // Issues of Turkish Philology. – 2025. – Т. 1, № 1. DOI: [10.5281/zenodo.15437929](https://doi.org/10.5281/zenodo.15437929)

LEMMATIZATION AS A TOOL OF NATURAL LANGUAGE PROCESSING: THEORETICAL AND APPLIED ASPECTS

Natavan Rahmanova

*Institute of Linguistics named after I. Nasimi, Azerbaijan National Academy of Sciences
Baku, Azerbaijan, natis-sdunina@mail.ru*

The article examines lemmatization as one of the key tools of natural language processing and morphological text analysis. The theoretical foundations of the concept of lemma, the specific features of the lemmatization process, and its differences from stemming are analyzed. Particular attention is paid to the application of lemmatization in agglutinative languages, especially in the Azerbaijani language, as well as to its role in automatic text analysis, information retrieval, and electronic lexicography. It is concluded that lemmatization is an important technology that ensures a more accurate interpretation of the semantic structure of texts in modern natural language processing systems.

Keywords: lemmatization; natural language processing; text analysis; lemma; stemming.