УДК 81’322:81’33
EDN AKDLRR
https://www.doi.org/10.33910/1992-6464-2025-217-302-311

ЛЕММАТИЗАЦИЯ МАЛОРЕСУРСНЫХ ЯЗЫКОВ В ДИАХРОНИЧЕСКОЙ ЛИНГВИСТИКЕ: ПРОБЛЕМЫ И РЕШЕНИЯ

Н. В. Дрожащих, Е. В. Ефимова

Аннотация

Введение. Статья посвящена проблематике лемматизации малоресурсных исторических языков в прикладной диахронической лингвистике. Нейросетевой (neural model) подход, используемый для лемматизации современных языков, для древних языков не применим из-за их морфологической сложности и ограниченности корпусных данных. Наиболее распространенным подходом к лемматизации малоресурсных исторических языков является словарно-ориентированный (dictionary-based) подход. В настоящее время наиболее доступный инструмент для лемматизации больших корпусов древнеанглийских текстов — модуль библиотеки Classical Language Toolkit (CLTK) — не позволяет осуществить корректную лемматизацию в силу ограниченности словаря лемм данной библиотеки. Цель исследования — компиляция словаря лемм древнеанглийского языка на основе данных краудсорсингового веб-ресурса — викисловаря (wiktionary) — для решения задач автоматической лемматизации.

Материалы и методы. Исследование проводится на материале корпуса аннотированных текстов-трибанков древнеанглийского языка (IX–XI вв.) и датасетов открытых лексикографических ресурсов (словарь лемм библиотеки CLTK и древнеанглийский сегмент викисловаря). Методологическая база исследования сочетает технологии корпусной и компьютерной лингвистики и электронной лексикографии для создания оригинального цифрового ресурса. Исследование включает три этапа: компиляция словаря древнеанглийских лемм, лемматизация древнеанглийских текстов, оценка точности (precision) и полноты (recall) словаря и качества лемматизации текста.

Результаты исследования. В ходе анализа лемматизаторов малоресурсных языков было выявлено, что оптимальным является подход с применением словаря лемм. Традиционные лексикографические источники, в частности древнеанглийский словарь Босворта-Толлера, содержат обширные словарные данные, однако эти данные не представлены в машиночитаемом формате. В качестве альтернативы мы предлагаем использовать викисловарь — открытый краудсорсинговый лексикографический ресурс, характеризующийся широким охватом и подробным описанием лексических единиц. В рамках исследования разработан алгоритм компиляции словаря лемм путем интеграции датасетов библиотеки CLTK и викисловаря. Полученный словарь лемм (11 451 уникальная лемма, 80 778 словоформ) показывает высокую степень точности и полноты, что подтверждает его применимость для лемматизации древнеанглийских текстов.

Заключение. По итогам исследования создан оригинальный цифровой ресурс — словарь лемм, а также программный код для автоматической лемматизации древнеанглийского языка. Разработанный нами код и полученный словарь лемм представляют собой эффективное решение задачи лемматизации малоресурсного исторического языка.

Ключевые слова: диахрония, малоресурсные исторические языки, древнеанглийский язык, лемматизация, лемматизатор, язык программирования Python, викисловарь

LEMMATIZATION OF LOW-RESOURCE LANGUAGES IN DIACHRONIC LINGUISTICS: PROBLEMS AND SOLUTIONS

N. V. Drozashchikh, E. V. Efimova

Abstract

Introduction. The article addresses the problem of lemmatizing low-resource historical languages within applied diachronic linguistics. Standard approaches widely used for modern texts — rule-based models and neural architectures — are inapplicable for ancient languages because of their morphological complexity and the scarcity of corpus data. As a result, dictionary-based lemmatization remains the most effective strategy. For Old English, the widely available Classical Language Toolkit (CLTK) offers only limited functionality, since its lemma dictionary provides insufficient coverage. The aim of this study is to compile an expanded lemma dictionary of Old English using data from the crowdsourced lexicographic resource Wiktionary, thereby enabling more accurate automatic lemmatization.

Materials and Methods. The study draws on the annotated treebank corpus of Old English texts (9th–11th centuries) and the datasets from the open lexicographic resources (the CLTK lemma list and the Old English segment of Wiktionary). The methodological framework combines methods of corpus and computational linguistics, and electronic lexicography. The research consists of three stages: compilation of an Old English lemma dictionary, lemmatization of Old English texts, and evaluation of the dictionary's precision and recall.

Results. The analysis of the lemmatizers for low-resource languages demonstrates that the dictionary-based approach proves to be more effective. Traditional lexicographic resources such as Bosworth–Toller remain valuable, but their lack of machine-readable formats limits direct application. As an alternative we propose Wiktionary — a free, crowdsourced lexicographic resource characterized by both broad lexical coverage and detailed descriptions of lemmas. Within the framework of this study, we developed a compiled lemma dictionary by integrating CLTK lemma list and Wiktionary datasets. The resulting lemma dictionary (11,451 unique lemmas, 80,778 wordforms) demonstrates a high degree of precision and recall, confirming its applicability for the lemmatization of Old English texts.

Conclusion. The study thus provides a new digital resource — an Old English lemma dictionary accompanied by Python code for fully automatic lemmatization. Together, these tools represent an effective solution to the lemmatization of a low-resource historical language and contribute to the development of computational diachronic linguistics.

Keywords: diachrony, low-resource historical languages, Old English, lemmatization, lemmatizer, Python programming language, Wiktionary

Введение

Настоящее исследование посвящено проблематике лемматизации малоресурсного древнеанглийского языка — важной задаче автоматической обработки естественного языка (Natural Language Processing — NLP). Лемматизация — это алгоритмический процесс приведения слов к словарной канонической форме — лемме. Данный процесс применяется для широкого спектра задач: подсчета частотности лемм, построения тематических моделей, стилометрического анализа, установления авторства текста, получения векторных представлений слов для анализа диахронических семантических изменений. В зависимости от масштаба задач и характера имеющихся ресурсов для автоматической лемматизации применяются различные подходы: словарно-ориентированный (dictionary-based), правило-ориентированный (rule-based), нейросетевой (neural model) и гибридный (hybrid).

Лемматизация в диахронической лингвистике представляет одну из самых сложных проблем NLP (Kanerva et al. 2021; Karimov et al. 2017; Saunack et al. 2021). Основные трудности связаны с ограниченностью

текстовых данных для создания аннотированных корпусов, богатой и сложной флективной морфологией языка, широким спектром семантических изменений в рамках самóго древнего языка, нестабильностью орфографии. Это осложняет автоматическую лемматизацию, основанную как на грамматических правилах, так и на нейронных моделях. В данных условиях словарно-ориентированный подход остается одним из наиболее эффективных способов лемматизации, требующим вместе с тем привлечения и расширения лексикографических ресурсов. Актуальность темы обусловлена необходимостью использования новых открытых лексикографических источников для решения прикладных задач диахронической лингвистики. Использование краудсорсинговых ресурсов, в частности викисловаря, позволяет расширить инструментарий автоматической обработки древних текстов.

Разработкой лемматизаторов древнеанглийского языка занимались ряд ученых (Fernández 2020; Percillier, Trips 2020; Rodríguez, Sáenz 2018; Torre 2022); отдельными исследователями/коллективами собраны корпусы древнеанглийских текстов (Bech, Eide 2014); для обработки текстовых корпусов использованы пайплайны Classical Language Toolkit (CLTK) для анализа древних текстов (Johnson et al. 2014–2021). Тем не менее отсутствует словарь лемм, который обеспечивал бы достаточный охват корпуса, единообразие и воспроизводимость результатов лемматизации древнеанглийских текстов.

Цель статьи — компиляция авторами настоящей статьи словаря лемм древнеанглийского языка на основе данных краудсорсингового веб-ресурса — викисловаря — для решения задач автоматической лемматизации. Задачи исследования: 1) аналитический обзор существующих подходов к лемматизации малоресурсных исторических языков; 2) извлечение лексикографической информации из викисловаря для построения словаря лемм; 3) интеграция базы лемм библиотеки CLTK и викисловаря для компиляции словаря лемм; 4) разработка программного кода для лемматизации древнеанглийского языка; 5) оценка точности и полноты итогового словаря и качества лемматизации размеченного корпуса древнеанглийских текстов проекта Information Structure and Word Order Change (ISWOC) in Germanic and Romance Languages Treebank (The ISWOC Project 2021).

Обзор литературы

Малоресурсными считаются малоизученные языки с небольшими корпусами данных, ограниченными или отсутствующими аннотированными данными, малым количеством носителей языка, а также языки, находящиеся под угрозой исчезновения, и языки с нестабильной орфографией на ранних стадиях своего развития (historical и/или non-standard languages) (Magueresse et al. 2020; Nigatu et al. 2024). В фокус внимания диахронической лингвистики попадают древнерусский (Eckhoff, Berdičevskis 2016), среднегреческий (Swaelens et al. 2024), латинский (Eger et al. 2015), нижнесаксонский/окситанский (Miletić, Siewert 2023), древне-/среднеанглийский (Fernández 2020; Percillier, Trips 2020; Rodríguez, Sáenz 2018; Torre 2022), древнефинский (Hämäläinen et al. 2021), средненидерландский/средневерхненемецкий (Manjavacas et al. 2019) и другие. Для лемматизации древних малоресурсных языков выделяют словарно-ориентированный, правило-ориентированный, нейросетевой и гибридный подходы. Для оценки работы существующих лемматизаторов ученые используют традиционные метрики: точность (accuracy; precision), охват (coverage), полнота (recall) и другие.

Словарно-ориентированный подход, основанный на заранее составленном словаре лемм, наиболее распространен для лемматизации флективных языков, характеризующихся широким спектром нестандартных морфологических форм и словообразовательных моделей. В исследовании (Dereza 2016) представлена методика компиляции словаря лемм для лемматизации древнеирландского языка, а также алгоритм предсказания лемм, отсутствующих в словаре, с использованием расстояния Дамерау — Левенштейна. В проекте

(Percillier, Trips 2020) предложен метод создания словаря лемм для текстов среднеанглийского периода с интеграцией лемм в существующие корпусные ресурсы.

Нейросетевой подход предполагает лемматизацию с применением предобученных нейронных сетей и традиционно используется для языков с достаточным количеством аннотированных корпусных данных. В исследовании (Hämäläinen et al. 2021) для лемматизации древнефинских текстов разработан нейронный пайплайн для орфографической нормализации и лемматизации, который демонстрирует высокую точность на неаннотированных данных. В работах (Gogoi, Baruah 2022; Saunack et al. 2021) применяется разновидность рекуррентных нейронных сетей, например Long Short Term Memory (LSTM)-модель для лемматизации индийских языков, и сделан вывод о зависимости точности лемматизации от объема обучающего корпуса.

При гибридном подходе лемматизация осуществляется с использованием словарей лемм / грамматических правил и проверки / коррекции результатов. В проекте Nerthus исследователи создают базу лемм для сильных глаголов древнеанглийского языка на основе доступных корпусных данных, учебных материалов и словарей, а лемматизация слабых глаголов осуществляется с использованием грамматических правил (Rodríguez, Sáenz 2018; Torre 2022). Полученные леммы интегрируются в синтаксически размеченные древнеанглийские корпусы (treebanks), доступ к которым осуществляется через специализированное программное приложение. В работе (Karimov et al. 2017) разработан инструмент для лемматизации текстов среднеанглийского периода, который включает словарь словоформ, частеречную разметку и вариативные орфографические формы.

Важной библиотекой, предоставляющей инструменты для обработки классических языков, является библиотека CLTK (Johnson et al. 2014–2021). Лемматизация в данной библиотеке реализована на базе словаря лемм / грамматических правил для морфологически сложных языков и предобученных нейронных моделей на аннотированных корпусах большего объема (Johnson et al. 2014–2021).

Рассмотрев существующие подходы и проанализировав работу лемматизаторов малоресурсных языков, мы пришли к следующим выводам. Правило-ориентированный и нейросетевой алгоритмы лемматизации, демонстрирующие высокую эффективность для современных языков, для малоресурсных исторических языков оказываются ограниченно применимыми в связи с отсутствием достаточного объема аннотированных данных, нестабильностью орфографии, высокой степенью вариативности форм. В настоящем исследовании мы предприняли попытку осуществить автоматическую лемматизацию древнеанглийского языка при помощи специально скомпилированного авторами статьи словаря лемм.

Викисловарь — открытый лексикографический ресурс

Автоматическая лемматизация текстов малоресурсных языков требует наличия словаря, сопоставляющего словоформы с их каноническими формами — леммами. Составление такого рода словарей представляет собой нетривиальную задачу. Традиционные словари, как правило, не предоставляют свои базы данных в электронном формате, доступном для свободного скачивания и автоматической обработки, что существенно ограничивает их применение в задачах автоматической обработки языка. В качестве альтернативы выступают краудсорсинговые веб-платформы, среди которых особое место занимает викисловарь (wiktionary). Подобные ресурсы становятся важным источником лексикографических данных. Растущий интерес научного сообщества к этим ресурсам подтверждается рядом работ, посвященных изучению возможностей применения викитехнологий, созданию на их основе датасетов, обоснованию генерации контента для параллельных словарей и разработке алгоритмов лемматизации для различных языков (Mausam et al. 2009; Navarro 2009; Hathout

2014; Liebeck, Conrad 2015; Bajčetić, Declerck 2022). Исследования демонстрируют сопоставимость результатов, полученных с использованием открытых краудсорсинговых ресурсов, с результатами, достигаемыми профессиональными лексикографами (Zesch, Gurevych 2009).

Викисловарь, запущенный для редактирования в 2004 году, представляет собой многоязычную веб-платформу для коллективного создания лексикографического контента (Wiktionary… 2025). Ресурс содержит более 1,4 миллиона словарных статей на 600 языках, включая информацию о фонетических, морфологических, синтаксических, семантических и этимологических свойствах лексических единиц, а также дефиниции и примеры употребления. Концепция викисловаря предполагает всестороннее описание лексики естественных и основных искусственных языков, имеющих письменность. Благодаря распределенной модели пользователи совместно формируют и редактируют словарные статьи.

Традиционные лексикографические источники древнеанглийского языка представлены такими основными ресурсами, как An Anglo-Saxon Dictionary (Bosworth 2014), The Dictionary of Old English Corpus (Healey et al. 2009), A Thesaurus of Old English (Roberts et al. 2017) и The Historical Thesaurus of English (Kay et al. 2025). Официальные количественные показатели о числе лемм в указанных источниках не опубликованы. Однако наблюдения над словарем An Anglo-Saxon Dictionary указывают на примерное количество 26 000 лемм (Martín Arista 2025: 5). В древнеанглийском фрагменте викисловаря число лемм составляет 56 047 вхождений. Таким образом, по количественным параметрам викисловарь может рассматриваться как ценный лексикографический источник.

Несмотря на широкое использование, викисловарь подвергается критике в научном сообществе относительно достоверности представленной информации из-за краудсорсингового характера наполнения. Между тем в отношении исторических языков, в частности древнеанглийского, следует отметить наличие редакторских инструкций в викисловаре, предписывающих привязку записей к засвидетельствованным источникам (Wiktionary… 2025). Более того, в (Kurmas 2010) приводится анализ грамматических данных из викисловаря, подтверждающий их высокую надежность. Для проверки корректности данных исследователь сопоставляет их с авторитетными лексикографическими источниками. Так, при анализе случайной выборки из 4748 слов для польского языка было исправлено лишь 2,75 % грамматических форм (131 единица), что свидетельствует о стабильности этих грамматических данных (Kurmas 2010).

Таким образом, неприменимость существующих традиционных лексикографических ресурсов для древнеанглийского языка в формате, пригодном для автоматической обработки, обусловливает целесообразность обращения к открытым краудсорсинговым источникам. Викисловарь представляет ценный ресурс для компиляции словаря лемм, необходимого для автоматической лемматизации древнеанглийских текстов.

Материал и методология исследования

Исследование проводится на базе древнеанглийского языка IX–XI веков. Древнеанглийский язык — это язык синтетического строя, отличающийся вариативностью грамматических формантов, в частности внешней и внутренней флексии (аблаут) (Hogg 1992: 31, 45, 151). Так, окончание *-on* в глаголе *cunnon* ‘можем’ является примером внешней флексии, а чередование *ō-ē* в глаголах *fōn* ‘ловить’ — *fēng* ‘поймал’ представляет собой пример аблаута. Особую группу составляют супплетивные глаголы, словоформы которых строятся на основе исторически разных корней. Например, словоизменительная парадигма древнеанглийского глагола *bēon* включает формы, образованные от трех индоевропейских корней: **bheuə-* ‘to be, exist, grow’ (‘быть, существовать, расти’); **es-* ‘to be’ (‘быть’); **wes-* ‘to stay, dwell, pass the night’

(‘останавливаться, обитать, ночевать’) (Watkins 1985, 8, 17, 78). Подобная сложность и вариативность морфологии древнеанглийского языка, а также нестабильность орфографии и ограниченность размеченных корпусных данных обусловливают преимущественное применение словарного подхода к лемматизации.

В качестве материала исследования используются: 1) датасеты открытых лексикографических ресурсов — база лемм библиотеки CLTK и база викисловаря (для первого этапа исследования); 2) трибанк проекта ISWOC (The ISWOC Project 2021), включающий пять текстов (28 300 словоформ): Ælfric’s Lives of Saints, Apollonius of Tyre, Anglo-Saxon Chronicles, Orosius, West-Saxon Gospels (для второго и третьего этапов).

Методология исследования сочетает технологии корпусной и компьютерной лингвистики и электронной лексикографии. На первом этапе — компиляция словаря древнеанглийских лемм применяются технологии парсинга, извлечения лексикографической информации, интеграции датасетов, дедубликации данных, унификации структуры словаря. Второй этап — лемматизация древнеанглийских текстов — использует технологии словарно-ориентированного подхода: точный поиск словоформы в предварительно скомпилированном словаре лемм, сопоставление словоформы и ее леммы и возврат леммы. На третьем этапе вычисляются метрики охвата, точности и полноты словаря и F-меры. Исследование реализовано на языке программирования Python. Для извлечения данных из архивного файла викисловаря и построения словаря лемм использовались стандартная библиотека Python и библиотеки zimply и pandas. Лемматизация, фильтрация стоп-слов, сопоставление лемм с корпусом ISWOC и расчет метрик качества выполнялись при помощи собственных классов (ISWOC, Lemmatizer). Программный код и скомпилированный словарь лемм представлены на веб-сервисе GitHub (Efimova 2025).

Результаты исследования

1. Компиляция словаря лемм

В рамках настоящего исследования составлен словарь лемм, предназначенный для задач автоматической лемматизации. Лексикографическая база сформирована на основе двух источников: библиотеки CLTK и викисловаря (Wiktionary… 2022). Файл с древнеанглийскими словоформами из библиотеки CLTK в формате txt доступен в открытом репозитории (CLTK... 2023). Для извлечения древнеанглийских лемм из базы данных викисловаря (Wikimedia Foundation… 2025) использован готовый пайплайн, реализованный на языке программирования Python, который позволил обработать полный архив англоязычного викисловаря в формате zim (UniMorph Project… 2022). В результате последовательного выполнения Python-скриптов на выходе сформирован корпус данных в формате txt, содержащий идентифицированные древнеанглийские леммы (существительные, прилагательные, глаголы) и их словоформы.

Для формирования итогового словаря лемм реализован программный модуль в виде Python-класса. Реализация модуля осуществляется поэтапно: исходные текстовые файлы преобразуются в tsv-формат (табличный формат для дальнейшей обработки); эти файлы соединяются, дубликаты устраняются; диакритики для долгих гласных и палатализованных согласных удаляются. Данная процедура унифицирует орфографию между текстовыми корпусами из разных источников.

В итоговом словаре каждая словоформа сопоставлена с соответствующей леммой. Например, лемме *bensian* соответствуют словоформы *bensa*, *bensast*, *bensaþ*, *bensian*, *bensia*, *bensiende*, *gebensod*. Итоговый словарь включает 11 451 уникальную лемму и 80 778 словоформ: 5878 лемм и 40 538 словоформ были получены из CLTK, а 7158 лемм и 50 542 словоформы — из базы данных викисловаря. Итоговый словарь является обширным систематизированным ресурсом,

предназначенным для автоматической обработки древнеанглийских текстов.

2. Лемматизация корпуса древнеанглийского языка

Лемматизация осуществляется на материале корпуса ISWOC с использованием скомпилированного нами словаря лемм. Для каждой словоформы в корпусе ISWOC приведена лемма, подтвержденная экспертной верификацией, например: *Mæg (mag) gehyran (gehyran) se (se) ðe (þe) wyle (willan) be (be) þam (se) halgan (halig) mædene (mægden) Eugenian (Eugenia) Philyppus (Philippus) dæhter (dohtor)* ‘Пусть услышит тот, кто желает о святой деве Евгении, дочери Филиппа’. Леммы расположены рядом с исходными словоформами и помещены в круглые скобки, например во фрагменте текста *mæg (mag)* единица *mæg* представляет собой исходную словоформу, а лексема в круглых скобках *(mag)* является леммой для данной словоформы. Для лемматизации используются исходные словоформы: *Mæg gehyran se ðe wyle be þam halgan mædene Eugenian Philyppus dæhter.*

Процесс лемматизации корпуса ISWOC начинается с инициализации лемматизатора, в ходе которой скомпилированный словарь преобразуется в структуру данных типа «словарь», где ключом выступает словоформа, а значением — ее каноническая лемма. Далее исходный текст подвергается токенизации, за которой следуют этап фильтрации стоп-слов и итеративный поиск леммы для каждой словоформы.

Результаты лемматизации сохраняются в двух форматах: в виде сплошного лемматизированного текста (*mag gehyran se þe none be se halig mægden Eugenia Philippus none*) и в виде кортежей, состоящих из словоформы и леммы, найденной в словаре (*mægden — mæden)*. Неидентифицированные словоформы обозначаются маркером ‘none’, например: *willan — none.*

Для древнеанглийского языка характерны ситуации орфографической вариативности и грамматической омонимии. Например, леммы *hi* и *hie* ‘они’ являются орфографическими вариантами; прилагательное *halig* ‘святой’ и существительное *halga* ‘святой человек’ — грамматическими омонимами. При обработке таких случаев указываются все возможные леммы, разделенные вертикальной чертой, например *hi|hie; halig|halga.*

3. Оценка точности и полноты словаря лемм и качества лемматизации

Оценка точности и полноты скомпилированного словаря лемм проводится путем сравнения результатов автоматической лемматизации с леммами, представленными в корпусе ISWOC. Для оценки качества используются стандартные метрики — охват, точность, полнота и F1-мера.

Количественный анализ охватывает 28 300 словоформ корпуса ISWOC, из которых 26 399 (93,28 %) были успешно сопоставлены с корректными леммами, найденными в словаре. При этом показатель охвата (доля слов корпуса, для которых предложена хотя бы одна лемма) составил 96,8 %. Показатель точности (часть предсказанных лемм, которые совпали с леммами корпуса ISWOC) равен 93,2 %. Полнота (доля всех корректных лемм из корпуса ISWOC) зафиксирована на уровне 92,1 %. F1-мера (сбалансированный показатель точности и полноты) оценивается в 92,6 %. Количество уникальных словоформ, для которых не удалось найти соответствующую лемму, составило 542 (1,92 %). Среди них 74 словоформы (13,65 %) — имена собственные, которые не всегда включаются в словарь лемм; часть словоформ — орфографические варианты лемм; для ряда единиц соответствующие леммы в словаре отсутствуют. Все эти случаи, помимо прочего, фиксируют ключевые проблемы лемматизации малоресурсных языков в диахронической лингвистике: ономастические пробелы, отсутствие нормализации орфографических вариантов, неполный словарный охват. В целом составленный словарь лемм демонстрирует достаточную полноту и точность, что подтверждает его применимость для задач лемматизации древнеанглийских текстов.

Заключение

В настоящем исследовании мы проанализировали работу существующих лемматизаторов древнеанглийского языка, протестировали викисловарь как достоверный ресурс для лемматизации древнеанглийских текстов (данный ресурс пополняет существующие базы лемм древнеанглийского языка, является удобным в обработке и применении), предложили свой программный код для лемматизации древнеанглийского языка, создали словарь (максимально полную базу лемм) и провели оценку точности и полноты полученного словаря. Программный код размещен в открытом доступе на веб-сервисе GitHub (Efimova 2025). Результаты исследования показали, что скомпилированный нами словарь лемм представляет собой эффективное решение проблемы лемматизации древнеанглийских текстов, позволяет с высокой степенью точности осуществлять их лемматизацию.

В ходе дальнейшего исследования планируется изучить возможности для решения проблем, поставленных в статье, а также изучить возможности и решения, предоставляемые большими языковыми моделями, для лемматизации малоресурсных языков.

SOURCES

Bech, K., Eide, K. (2014) *The ISWOC corpus. Department of Literature, Area Studies and European Languages, University of Oslo*. [Online]. Available at: http://iswoc.github.io (accessed 23.05.2025). (In English)

CLTK. (2023) *Old English Lemmas: oe.lemmas*. [Online]. Available at: https://github.com/cltk/ang_models_cltk/blob/master/data/oe.lemmas (accessed 23.05.2025). (In English)

Efimova, E. V. (2025) *OE_Lemmatization*. [Online]. Available at: https://github.com/webnora/OE_Lemmatization (accessed 23.05.2025). (In English)

Johnson, K. P., Burns, P., Stewart, J., Cook, T. (2014–2021)*CLTK*: *The Classical Language Toolkit*. [Online]. Available at: https://github.com/cltk/cltk (accessed 05.03.2025). (In English)

The ISWOC Project. (2021) *The ISWOC Treebank*. [Online]. Available at: https://dev.syntacticus.org/iswoc.html#downloads (accessed 23.05.2025). (In English)

UniMorph Project. (2022) *Universal Morphology (UniMorph)*. [Online]. Available at: https://github.com/unimorph (accessed 23.05.2025). (In English)

Wikimedia Foundation. (2025) *Index of /enwiktionary/*. [Online]. Available at: https://dumps.wikimedia.org/enwiktionary/ (accessed 23.05.2025). (In English)

Category: Old English lemmas. (2022). Wiktionary. [Online]. Available at: https://en.wiktionary.org/wiki/Category:Old_English_lemmas (accessed 23.05.2025). (In English)

DICTIONARIES

Bosworth, J. (2014) *An anglo-saxon dictionary online*. [Online]. Available at: https://bosworthtoller.com/ (accessed 23.05.2025). (In English)

Healey, A., diPaolo, A., Holland, J. et al. (2009) *The dictionary of Old English corpus in electronic form, TEI-P5 conformant version* [CD-ROM]. Toronto: Dictionary of Old English Project, University of Toronto. (In English)

Kay, C., Alexander, M., Dallachy, F. et al. (eds.) (2025) *The historical thesaurus of English*. 2nd ed., version 5.0. Glasgow: University of Glasgow Publ. [Online]. Available at: https://ht.ac.uk/ (accessed 23.05.2025). (In English)

Roberts, J., Kay, C., Grundy, L. (2017) *A Thesaurus of Old English*. Glasgow: University of Glasgow Publ. [Online]. Available at: http://oldenglishthesaurus.arts.gla.ac.uk/ (accessed 23.05.2025). (In English)

Watkins, C. (1985) *The American heritage dictionary of Indo-European roots*. Boston: Houghton Mifflin Publ., 149 p. (In English)

Wiktionary. (2025) [Online]. Available at: https://en.wiktionary.org/wiki/ (accessed 23.05.2025). (In English)

REFERENCES

Bajčetić, L., Declerck, T. (2022) Using Wiktionary to create specialized lexical resources and datasets. In: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022) June 20–25, 2022*. Marseille: European Language Resources Association Publ., pp. 3457–3460. (In English)

Dereza, O. (2016) Building a dictionary-based lemmatizer for Old Irish. In: *Actes de la conférence conjointe JEP-TALN-RECITAL*, *vol. 6: Celtic Language Technology Workshop*. Paris: AFCP — ATALA Publ., pp. 12–17. (In English)

Eckhoff, H., Berdičevskis, A. (2016) Automatic parsing as an efficient pre-annotation tool for historical texts. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) December 11–17 2016*. Osaka: The COLING 2016 Organizing Committee Publ., pp. 62–70. (In English)

Eger, S., vor der Brück, T., Mehler, A. (2015) Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization models. In: *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities July 30, 2015*. Beijing: Association for Computational Linguistics Publ., pp. 105–113. https://doi.org/10.18653/v1/W15-3716 (In English)

Fernández, L. G. (2020) Sources and steps of corpus lemmatization. Old English anomalous verbs. *Revista Española de Lingüística Aplicada — Spanish Journal of Applied Linguistics*, vol. 33, no. 2, pp. 416–442. https://doi.org/10.1075/resla.18024.gar (In English)

Gogoi, A., Baruah, N. A (2022) A lemmatizer for low-resource languages: WSD and Its role in the Assamese language. *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 4, article 74. https://doi.org/10.1145/3502157 (In English)

Hämäläinen, M., Partanen, N., Alnajjar, K. (2021) Lemmatization of historical old literary Finnish texts in modern orthography. In: *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), 28 juin au 2 juillet 2021 [Proceedings of the 28th Conference on Natural Language Processing (TALN), June 28 to July 2, 2021]*. Lille: ATALA Publ., pp. 189–198. https://doi.org/10.48550/arXiv.2107.03266 (In English)

Hathout, N., Sajous, F., Calderone, B. (2014) Acquisition and enrichment of morphological and morphosemantic knowledge from the French Wiktionary. In: *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing August 24 2014*. Dublin: Association for Computational Linguistics and Dublin City University Publ., pp. 65–74. https://doi.org/10.3115/v1/W14-5809 (In English)

Hogg, R. M. (ed.). (1992) *The Cambridge History of the English Language*. *Vol. 1: The Beginning to 1066.* Cambridge: Cambridge University Press, 613 p. (In English)

Kanerva, J., Ginter, F., Salakoski, T. (2021) Universal lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, vol. 27, no. 5, pp. 545–574. https://doi.org/10.1017/S1351324920000224 (In English)

Karimov, R., Samkova, M., Nikitina, S., Akinin, A. (2017) Using a hybrid algorithm for lemmatization of a diachronic corpus. In: *Proceedings of the Workshop on Computational Linguistics and Language Science, 26 April 2016 CLLS 2016*. Vol. 1886, pp. 1–8. [Online]. Available at: https://ceur-ws.org/Vol-1886/ (accessed 05.03.2025). (In English)

Liebeck, M., Conrad, S. (2015) iwnlp: inverse wiktionary for natural language processing. In: *Proceedings of the 53rd Annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, vol. 2: Short Papers*. Beijing: Association for Computational Linguistics Publ., pp. 414–418. https://doi.org/10.3115/v1/P15-2068 (In English)

Magueresse, A., Carles, V., Heetderks, E. (2020) *Low-resource languages: A Review of past work and future challenges*. [Online]. Available at: https://arxiv.org/pdf/2006.07264 (accessed 05.03.2025). (In English)

Manjavacas, E., Kádár, Á., Kestemont, M. (2019) Improving lemmatization of non-standard languages with joint learning. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, vol. 1 (Long and Short Papers*. Minneapolis: Association for Computational Linguistics Publ., pp. 1493–1503. https://doi.org/10.18653/v1/N19-1153 (In English)

Martín Arista, J. (2025) The Computational Study of Old English. *Encyclopedia*, vol. 5, no. 3, article 137. https://doi.org/10.3390/encyclopedia5030137 (In English)

Mausam, Soderland, S., Etzioni, O. et al. (2009) Compiling a massive, multilingual dictionary via probabilistic inference. In: *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, vol. 1*. Suntec: Association for Computational Linguistics Publ., pp. 262–270. (In English)

Miletić, A., Siewert, J. (2023) Lemmatization experiments on two low-resourced languages: Low Saxon and Occitan. In: *Tenth workshop on NLP for similar languages, varieties and dialects (VarDial 2023) May 5, 2023*. Dubrovnik: Association for Computational Linguistics Publ., pp. 163–173. https://doi.org/10.18653/v1/2023.vardial-1.17 (In English)

Navarro, E., Sajous, F., Gaume, B. et al. (2009) Wiktionary for Natural language processing: Methodology and limitations. In: *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources (People's Web)*. Suntec: Association for Computational Linguistics Publ., pp. 19–27. (In English)

Nigatu, H. H., Tonja, A. L., Rosman, B. et al. (2024) The Zeno's paradox of "low-resource" languages. In: *Proceedings of the 2024 conference on Empirical methods in natural language processing*. Miami: Association for Computational Linguistics Publ., pp. 17753–17774. https://doi.org/10.18653/v1/2024.emnlp-main.983 (In English)

Percillier, M., Trips, C. (2020) Lemmatising verbs in middle English corpora: The benefit of enriching the Penn-Helsinki parsed corpus of middle English 2 (PPCME2), the parsed corpus of middle English Poetry (PCMEP), and a parsed linguistic atlas of early middle English (PLAEME). In: *Proceedings of the 12th language resources and evaluation conference (LREC 2020) 11–16 May, 2020*. Marseille: European Language Resources Association Publ., pp. 7170–7178. (In English)

Sáenz, M. T., Rodríguez, D. M. (2018) A semiautomatic lemmatisation procedure for treebanks. Old English strong and weak verbs. In: *Proceedings of the 16th international workshop on treebanks and linguistic theories (TLT16) January 23–24, 2018*. Prague: [s. n.], pp. 88–94. (In English)

Saunack, K., Saurav, K., Bhattacharyya, P. (2021) How low is too low? A monolingual take on lemmatisation in Indian languages. In: *Proceedings of the 2021 Conference of the North American chapter of the association for computational linguistics: Human language technologies*. [Online]. Available at: https://doi.org/10.18653/v1/2021.naacl-main.322 (accessed 05.03.2025). (In English)

Swaelens, C., Singh, P., Vos, I. De, Lefever, E. (2024) Lemmatisation of medieval Greek: Against the limits of transformers' capabilities? In: *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024) 20–25 May, 2024*. Torino: ELRA Publ., pp. 10293–10302. (In English)

Torre, A. R. (2022) Automatic lemmatization of old English class iii strong verbs (L-Y) with ALOEV3. *Journal of English Studies*, vol. 20, pp. 237–266. https://doi.org/10.18172/jes.5324 (In English)

Wiktionary. (2025) Wikipedia. [Online]. Available at: https://en.wikipedia.org/wiki/Wiktionary (accessed 14.08.2025). (In English)

Wiktionary: Old English entry guidelines. (2025) *Wiktionary*. [Online]. Available at: https://en.wiktionary.org/wiki/Wiktionary%3AOld_English_entry_guidelines (accessed 21.08.2025). (In English)

Zesch, T., Gurevych, I. (2009) Wisdom of crowds versus wisdom of linguists — measuring the semantic relatedness of words. *Journal of Natural Language Engineering*, vol. 16, no. 1, pp. 25–59. https://doi.org/10.1017/S1351324909990167 (In English)

СВЕДЕНИЯ ОБ АВТОРАХ

ДРОЖАЩИХ Наталия Владимировна — *Nataliia V. Drozhashchikh*
Тюменский государственный университет, Тюмень, Россия.
Tyumen State University, Tyumen, Russia.
SPIN-код: 1215-4370; Scopus Author ID: 57219205143; ResearcherID: G-3203-2017; ORCID: 0000-0002-5910-2402, e-mail: n.v.drozhashhikh@utmn.ru
Доктор филологических наук, доцент, профессор кафедры прикладной и теоретической лингвистики.

ЕФИМОВА Елена Викторовна — *Elena V. Efimova*
Тюменский государственный университет, Тюмень, Россия.
Tyumen State University, Tyumen, Russia.
ORCID: 0000-0002-6584-965X, e-mail: e.v.efimova@utmn.ru
Старший преподаватель кафедры прикладной и теоретической лингвистики.

Поступила в редакцию: 15 мая 2025.
Прошла рецензирование: 23 июля 2025.
Принята к печати: 30 сентября 2025.