Министерство науки и высшего образования Российской Федерации
Федеральное государственное образовательное учреждение высшего образования
«Челябинский государственный университет»

На правах рукописи

Каримов Рауль Дамирович

АВТОМАТИЗАЦИЯ ПРОЦЕССА РАЗМЕТКИ ДИАХРОНИЧЕСКОГО КОРПУСА (НА МАТЕРИАЛЕ СРЕДНЕАНГЛИЙСКОГО И ДРЕВНЕСКАНДИНАВСКОГО ЯЗЫКОВ)

Специальность 10.02.20 — Сравнительно-историческое, типологическое и сопоставительное языкознание
10.02.21 — Прикладная и математическая лингвистика

Диссертация на соискание ученой степени
кандидата филологических наук

Научный руководитель:
доктор филологических наук, профессор
Л.А. Нефедова

Челябинск
2020

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Сравнительно-историческое языкознание, зародившееся как отрасль науки в первой половине XIX века, на сегодняшний день является одним из тех разделов лингвистики, которые в наибольшей мере полагаются на применение корпуса – специально размеченного, машино- и человеко-читаемого массива связных текстовых данных, предназначенного для конкретных целей и исследовательских задач. Об этом свидетельствует как востребованность и академическая поддержка отдельных проектов по созданию диахронического, или исторического, корпуса (например, Linguistic Atlas of Early Middle English, Helsinki Corpus of English Texts, Medieval Nordic Text Archive, Corpus of Early Middle English Poetry, каждый из которых создавался командами экспертов в языкознании и ИТ), так и популярность диахронически размеченных корпусов на платформах общего доступа, например, SketchEngine, где именно Time-Stamped JSI Web-Corpus of English является крупнейшим корпусом:

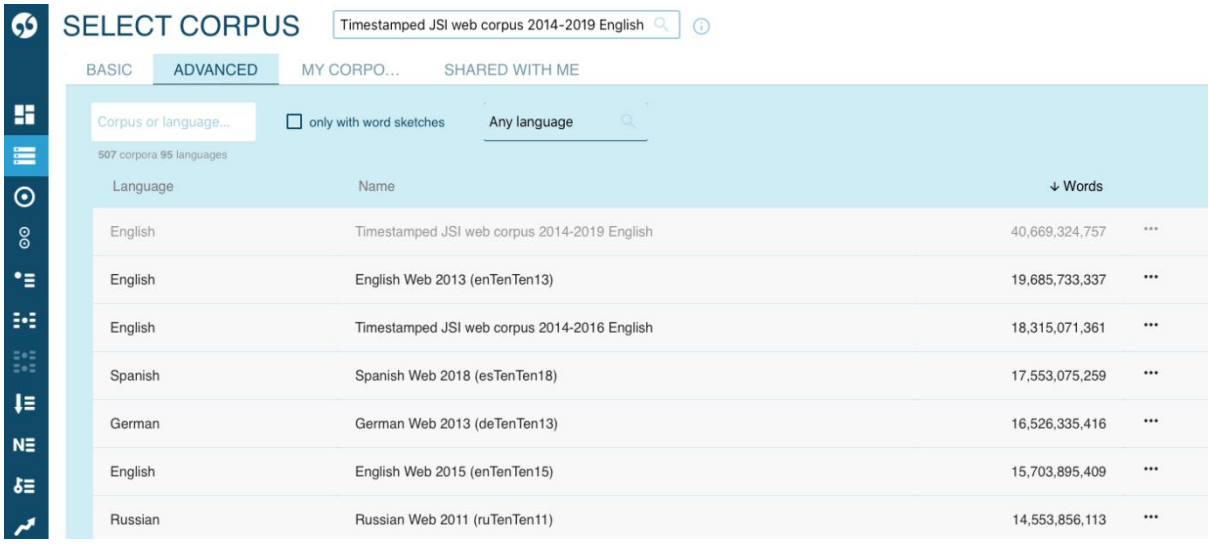


Рис. 1. Перечень корпусов на платформе SketchEngine с ранжированием по объему

Применение корпуса в историческом языкознании может относиться как к дескриптивным исследованиям, например, сопоставительному и сравнительному анализу развития концепта семьи в германских и славянских языках, так и к изысканиям, полагающимся на комплексный математический

аппарат, например, в глоттохронологии. Независимо от характера исследования, обращающегося к корпусу, в последнем необходима многоуровневая разметка: морфологическая, синтаксическая, семантическая, супрасегментальная, мета- и экстралингвистическая. Внесение подобной разметки вручную требует труда нескольких квалифицированных лингвистов при составлении даже относительно небольших корпусов, в связи с чем автоматизация данного процесса, по крайней мере, частичная, с помощью статистических методов и программных средств ЭВМ, является широко исследуемой проблемой в междисциплинарной области, известной как компьютерная лингвистика.

Еще на заре структуралистского подхода в лингвистике – в конце XIX и начале XX вв. – было отмечено, что распределение слов по частям речи в языке подчиняется определенным статистическим закономерностям, благодаря чему возникли так называемые скрытые марковские модели. Такие модели, названные по имени российского математика А. А. Маркова, впервые их сформулировавшего, описывают процессы смены состояний, при которых следующие за определенным моментом времени *t* состояния не зависят от предшествующих) — то есть представляют собой математическое предиктивное описание синтаксических последовательностей на основе вероятностного распределения. Дальнейшее развитие математического метода привело к возникновению вскоре после Второй Мировой войны так называемой «машины Тьюринга» — абстрактной модели (конечного автомата), на основе которой построены принципы компьютерных алгоритмов. Практически одновременно с данной машиной были сформулированы идеи Хомского о порождающих грамматиках – формальных определениях, использующих комбинации терминальных и нетерминальных символов для описания языковой структуры. Так появилось направление, впоследствии названное «вычислительной», или «компьютерной», лингвистикой (англ. computational linguistics). В своем развитии компьютерная лингвистика значительно опередила любые другие области филологии, при этом принадлежа скорее к междисциплинарным областям на стыке гуманитарного и математического подходов; появление

концепции лингвистического корпуса в 70-х гг. XX века поспособствовало дальнейшему распространению и прогрессу вычислительных методов в языкознании, развитию глоттохронологии, машинного перевода, автоматического реферирования текстов, анализа и синтеза текста, распознавания речи, других приложений. Исследования в области компьютерной лингвистики остаются актуальными и по сей день.

Одним из наиболее популярных направлений языкознания является автоматическое аннотирование корпуса текстов, что обусловлено, с одной стороны, большой трудоемкостью ручной работы по созданию разметки, с другой – широкими просторами для развития методов искусственного интеллекта применительно к анализу текста. Задача автоматического аннотирования, в частности, указания частей речи (англ. PoS-tagging) до сих пор не решена до конца даже для наиболее распространенных языков, таких, как английский: в то время как нормализованный текст, например, корпус газетных и журнальных публикаций, легко поддается такому аннотированию даже с применением классических алгоритмов (например, TreeTagger), орфографически или грамматически девиантный текст (например, корпус комментариев в социальных сетях) может вызывать большое количество ошибок аннотирования. Для нас в связи с этим представляет интерес эффективность реализации подобных алгоритмов применительно к историческим языкам (англ. historical language, русскоязычный термин «древний язык» не используем, т.к. не каждый исторический язык является древним. Так, в диахронической лингвистике принято различать древний и среднеанглийский язык, но оба являются историческими), в частности, германским, специфика которых заключается в том числе в недостаточной нормализации письменной формы. Этим обусловлена **актуальность** настоящего исследования – она состоит в нахождении компьютерного алгоритма, позволяющего в значительной степени или полностью автоматизировать процесс размечивания диахронического корпуса, что упростит задачу построения таких корпусов в будущем и позволит объяснить многие моменты в истории становления языка.

Объектом исследования стали среднеанглийский и древнескандинавский языки, в частности, их графическая реализация в форме предварительно размеченного малого корпуса текстов, а **предметом** – частеречно-морфографемические особенности изучаемого периода становления древних языков, а также возможности их анализа с использованием программных средств ЭВМ в целях дальнейшей автоматизации ведения частеречной разметки в историческом тексте.

Материалом исследования послужили два корпуса: Лингвистический атлас раннего среднеанглийского языка (Linguistic Atlas of Early Middle English, LAEME) – маркированный диахронический корпус, содержащий примерно 167 текстов периода с 1150 по 1325 гг. общим объемом около 816 тысяч слов; и Архив средневековых скандинавских текстов (Medieval Nordic Text Archive, Menota), содержащий около 2 миллионов слов исторических документов на древнескандинавском языке в различных вариантах записи (дипломатическая, факсимильная и нормализованная запись). Первый корпус имеется в открытом доступе на платформе GitHub; доступ ко второму был получен в рамках обучения в Бергенском университете.

Цель настоящего исследования состоит в том, чтобы выявить особенности частеречной разметки диахронического корпуса на материале древних языков, проанализировать их влияние на качественный результат применения компьютерных средств автоматизации такой разметки, одновременно обеспечив пригодность таких средств к применению лингвистами без существенной информационно-технологической подготовки.

Гипотеза исследования состоит в том, что морфографемическое кодирование текста (векторная репрезентация, обеспечивающая уникальную и однозначную идентификацию частеречной принадлежности лексемы по ее морфологически обусловленной графической реализации) позволяет достичь адекватной машиночитаемости среднеанглийского и древнескандинавского текста в контексте их уникальной исторически сложившейся грамматической, орфографической, фонетической специфики, в то время как обращение к

ансамбль-алгоритмам может обеспечить высокоэффективную автоматизацию процесса частеречной разметки при обучении с учителем на малой выборке.

Обозначенная цель и выдвинутая гипотеза обусловливают следующие **задачи**:

1) выявить и сопоставить особенности частеречной разметки, характерной для корпусов как среднеанглийского, так и древнескандинавского языков;
2) проанализировать рассматриваемый языковой материал в диахроническом срезе при помощи методов машинного обучения, как общих, так и созданных непосредственно в целях аннотирования текстов;
3) разработать алгоритм комбинирования различных методов обучения, оптимальный для исследования древних языков;
4) экспериментально проверить на материале среднеанглийского и древнескандинавского языков созданный алгоритм.

Для решения поставленных задач применяется совокупность различных **методов**, как общенаучных, так и специализированных аналитических и прикладных. Все применимые методы можно разделить на следующие категории в зависимости от целей их приложения:

a) сравнительно-исторический метод;

b) методы машинного обучения, не являющиеся собственно лингвистическими: теорема Байеса, k ближайших соседей, метод опорных векторов, модель случайного леса, линейная регрессия, многослойный перцептрон;

c) методы ансамблирования, позволяющие сочетать применение различных техник машинного обучения в рамках работы одного алгоритма;

d) компьютерно-лингвистические методы векторизации лексического материала: метод мешка слов, метод N-грамм, стеммирование, векторизация по скользящему среднему;

e) метод эксперимента, основанного на многократной перекрестной проверке точности и полноты машинного обучения;

f) гипотетико-дедуктивный метод, используемый для общей структуризации и интерпретации полученных результатов.

Теоретико-методологической основой работы послужили исследования отечественных и зарубежных ученых, посвященные историческим формам английского и скандинавских языков (М.И. Стеблин-Каменский, Б.А. Ильиш, О.Э. Хеуген, К. Далтон-Паффер, К. Бруннер и Г. Джонсон, Г. Брэдли, С. Хоробин и Дж. Смит, Р. Хогг, У.В. Ральф и М.А. Элиотт, Я.Т. Форлун, К. Фоссум и Э. Углан, Р. Мактюрк, О. Бэндл, Х. О'Донаф, Дж. Клаксон, Г. Свит, Т. Мустаноя, Э. ван Хельдрен, В.Д. Аракин, Э. Кенихь и Й. ван дер Аувера, Л. Кэмпбелл, У. Стаббс и Х.В.К. Дэвис, К.М. Милуорд и М. Хэйз, Р. Бенедиктсон и Э. Мундаль и др.), а также теории и практике машинного обучения в лингвистике (Д. Джурафски, Т. Мун, Дж. Болдридж, Х. Шютце и К.Д. Маннинг, М. Минский, Б. Галицкий, Дж. Челано, К. Якобини, С. Хельгадоттир, Н. Хомски и Т. Виноград, К. Шэннон и др.), общим вопросам информационных технологий и искусственного интеллекта (А. Тьюринг и У. Бледсоу, С. Маккаллох и Дж.В. Бакус, Ф. Мостеллер и Р.Э. Лонгакр, и др.).

Научная новизна работы состоит в том, что в ней впервые:

• производится попытка сравнить характер частеречной разметки корпуса на среднеанглийском и древнескандинавском языках;

• комбинируются и сопоставляются различные алгоритмы, основанные на машинном обучении, применительно к существующему размеченному корпусному материалу в малых фрагментах;

• выявляются особенности диахронического развития среднеанглийского и древнескандинавского языков, оказавшие влияние на качественные характеристики текстового корпуса и значимые в контексте машинного обучения на его материале;

• используется посимвольная репрезентация, позволяющая адекватно векторизовать входные текстовые данные в относительно малой размерности;

• описывается упрощенный метод аннотирования на основе вводимых пользователем данных (т.н. «обучение с учителем», англ. supervised learning);

Теоретическая значимость работы состоит в сопоставительном и сравнительно-историческом выявлении диахронических эволюционных черт анализируемых языков. Применение эффективных методик и алгоритмов машинного обучения обусловливает **практическую значимость**, которая заключается в синтезе скорректированного в контексте выявленных диахронических черт машинно-обученной модели разметки, обеспечивающей простоту и эффективность ведения малого исторического корпуса одним исследователем в рамках собственных прикладных задач; появление такой модели позволит, с одной стороны, значительно облегчить подготовительную работу в рамках проектов, связанных с историческими языками, с другой — даст лингвистам, не имеющим основательной подготовки в области методов машинного обучения, возможность успешно применять такие методы в своих исследованиях.

На защиту выносятся **следующие положения**:

1. Среднеанглийский и древнескандинавский языки обладают рядом особенностей, преимущественно относящихся к орфографии, которые требуют специального подхода при составлении алгоритма автоматического аннотирования в рамках составления диахронического корпуса.

2. Орфографические особенности среднеанглийского и древнескандинавского языков, их морфологическая и фонетическая специфика оказывают влияние на качественные характеристики автоматического частеречного аннотирования корпуса текстов на древних языках.

3. Учесть эволюционную особенность становления анализируемых древних языков можно обобщенным методом, базирующимся на морфографемическом векторном кодировании текста.

4. Наибольшей эффективности при исследовании изучения генетически связанных родственных языков и установления соотношения между родственными языками и описания их эволюции во времени и пространстве

можно добиться машинным обучением путем ансамблирования, то есть комбинирования вывода различных алгоритмов, в том числе с наложением весовых коэффициентов, что обусловлено контекст-специфичной предиктивной силой разных алгоритмов.

5. Разработанный и предложенный инструмент значительно упрощает работу исследователя по составлению исторического корпуса, в особенности в условиях недостаточности трудового ресурса.

Достоверность и обоснованность полученных в рамках исследования практических результатов обеспечиваются применением многопроходной перекрестной проверки, снижающей вероятность влияния случайных событий на точность и полноту выводов алгоритма машинного обучения; использованием метода определения статистической значимости, в частности, парного критерия Стьюдента.

Структура и объем диссертационного исследования обусловлены поставленной целью и решаемыми задачами. Работа состоит из введения, двух глав, разделенных на параграфы, заключения, списка литературы, списка табличных и графических вложений.

Во **Введении** обосновывается актуальность темы исследования, раскрывается степень её разработанности; формулируются объект, предмет исследования, его цель и задачи; описывается материал исследования; перечисляются методы, использованные при выполнении работы; выдвигается гипотеза, приводятся положения, выносимые на защиту; определяются научная новизна, теоретическая и практическая значимость исследования; указываются сведения о степени достоверности и апробации результатов исследования.

В **первой** главе дается характеристика исторического текста на среднеанглийском и древнескандинавском языках, анализируется их проблематика с точки зрения задач машинного обучения; излагаются принципы работы общих методов машинного обучения, методов, разработанных и применяемых непосредственно в рамках решения задач аннотирования корпусов, а также различных способов машиночитаемого представления текста.

Во **второй** главе проводится сравнительный анализ двух изучаемых древних языков, подробно описывается материал и его подготовка, разработанный способ векторизации текста в целях обеспечения его машиночитаемости; представлен отчет об экспериментальной проверке эффективности предложенного алгоритма и метода репрезентации, а также эксплуатационной пригодности его реализации в виде программы: оцениваются точность и полнота разметки, собираются и анализируются отзывы пользовательской группы, выполняется расчет эффективности размечивания с помощью программы по сравнению с полностью ручной разметкой; глава также содержит обсуждение результатов каждого из проведенных экспериментов.

Диссертационное исследование завершается **заключением**, в котором описываются проделанная работа и достигнутые результаты с точки зрения выдвинутой гипотезы, приводятся размышления о возможностях дальнейшего направления данных научных изысканий.

После заключения приводится **список литературы** (170 единиц).

Диссертационное исследование выполнено в рамках содержания и областей знаний, определяемых паспортами специальностей 10.02.20 «Сравнительно-историческое, типологическое и сопоставительное языкознание»:

- разработка и развитие языковедческой теории и методологии на основе изучения генетически связанных родственных языков и установления соотношения между родственными языками и описания их эволюции во времени и пространстве;
- изучение структурных и функциональных свойств языков независимо от характера генетических отношений между ними;
- выявление различий (контрастности) между двумя сравниваемыми языками;
- сравнение и сопоставление языков в диахронии и синхронии;

и 10.02.21 «Прикладная и математическая лингвистика»:

- развитие теории и методологии прикладной и математической лингвистики в диахронии и синхронии;
- статистико-комбинаторное моделирование;
- вероятностные и статистические модели языка и речи;
- лингвистическая статистика (вычислительная, квантитативная, количественная, статистическая лингвистика);
- компьютерная лексикология и лексикография.

К настоящему моменту работа **апробирована** представлением ее результатов на 13 конференциях, форумах и семинарах, включая доклады на семинаре «Computational Linguistics and Language Studies», г. Москва (2016 г.), Международных научно-практических конференциях студентов, аспирантов и молодых ученых «Ломоносов», г. Москва (2018, 2019 гг.), Международной научно-практической конференции «Индустрия перевода», г. Пермь (2017 и 2019 гг.), Международной конференции «Theoretical and Applied Linguistics», г. Белосток, Польша (2017 г.), Международной конференции «English Language and Anglophone Literatures Today», г. Нови-Сад, Сербия (2017 г.), Международной конференции «Analysis of Images, Social Networks and Texts (AIST)», г. Москва (2018 г.), Международной научной конференции «Слово, высказывание, текст в лингвистическом, культурологическом и прагматическом аспектах», г. Челябинск (2018 г.) и на Международной конференции «Актуальные проблемы лингвистики», г. Костанай (2018 г.); работа дважды представлена к участию в конкурсе «Наука будущего — наука молодых» в 2017 и 2018 гг., оба раза становилась финалистом конкурса.

Результаты работы **опубликованы** в 2 статьях в периодических изданиях из списка ВАК, 2 статьях в составе сборников, индексируемых базой Scopus, 1 статье в европейском журнале Белостоцкого университета, 2 статьях в сборниках, индексируемых РИНЦ, и 4 тезисных статьях; получен патент на разработанное при участии сотрудников математического факультета ФГБОУ ВО «ЧелГУ» программное обеспечение для ЭВМ. Личный вклад автора работы заключался в подборке методов машинного обучения и создании на их основе

комбинированного алгоритма, подходящего для анализа текстов на среднеанглийском и древнескандинавском языках, а также в подготовке обучающих и проверочных выборок, проведении экспериментов.

Кроме того, в рамках работы получен грант по программе российско-норвежских учебных грантов, благодаря которому автор диссертации в 2017–2018 учебном году в течение 2 семестров обучался и вел исследовательскую работу на гуманитарном факультете Университета г. Берген, Норвегия.

Автор диссертации выражает благодарность Акинину А.А. и Якимцу Д.Н. за помощь в работе со средствами машинного обучения.

ГЛАВА I. ТЕОРЕТИКО-МЕТОДОЛОГИЧЕСКАЯ ПРОБЛЕМАТИКА ЧАСТЕРЕЧНОЙ РАЗМЕТКИ ДИАХРОНИЧЕСКОГО КОРПУСА

В последние десятилетия одним из важнейших инструментов лингвистических исследований становится т.н. корпус текстов – репрезентативный и прагматически ориентированный, т.е. создаваемый для определенных целей массив репрезентативных и размеченных текстовых данных [Захаров, 2016, с. 141]. Корпус, согласно этому определению, несет в себе разметку: машиночитаемый код, в котором указываются морфологические, синтаксические, семантические и экстралингвистические свойства приводимого текста и/или отдельных его компонентов (лексем). Разновидность корпуса и тип разметки обусловливаются целями исследования; так, в переводоведении широко применимы параллельные корпусы [Краснопеева, 2015], в исследовании семантики и статистическом анализе словоупотреблений – общие корпусы, а в диахронической лингвистике – исторические, например, Helsinki Corpus of English Texts. Корпус также служит незаменимым источником информации в компьютерной лингвистике, обеспечивая исследователей большим объемом маркированных (labeled) данных для обучения компьютерных алгоритмов, например, в целях генерации пользовательских моделей для движка нейросетевого перевода от Microsoft Corporation, позволяющего осуществлять машинный перевод с соблюдением стилистики предложенных к обучению исходных текстов [Что такое Custom Translator? - Azure Cognitive Services | Microsoft Docs]; последняя задача, в частности, в значительной мере опирается на параллельные корпусы [Заковоротная, Северина, 2018]. Корпусы с обширными метаданными, например, многонациональные корпусы текстов СМИ с указанием года и страны происхождения текста (доступны в том числе на бесплатной платформе SketchEngine) обеспечивают возможность лингвокультурологического анализа реализации тех или иных концептов. Наконец, литературоведческий корпус находит применение в исследованиях стилеметрии за счет вывода статистических параметров авторской лексики и

синтаксиса [Родионова, 2008]. Таким образом, практически все сферы современной науки о языке: от переводоведения до исторической лингвистики, от когнитивистики до литературоведения – могут тем или иным образом задействовать корпус текстов как исследовательский инструмент.

Возможно, наиболее широко такой инструмент используется в изучении именно истории языка. В той или иной форме диахронический корпус применим глоттохронологией в целях изучения эволюции словарного состава [Арапов, Херц, 1973] и выделения ядерных слов, формирующих ключевой понятийный аппарат носителей того или иного языка; сопоставительно-сравнительным языкознанием в целях выявления, хронологизации и детального описания различных историко-грамматических процессов, например, тематизации презенса сильного глагола в сопоставлении нескольких языков [Николаева, 2003]; в попытках установить хронологию тех или иных процессов фонетической эволюции, например, развития гармонии гласных [Hagland, 1978]; в рамках исследования исторической литературы в попытке найти отражение того или иного аспекта; социологами в целях установления и анализа проявлений того или иного общественно-психологического явления в историческом обществе, например, мужской гомосексуальности среди викингов [Gade, 1986]; в междисциплинарных исследованиях на стыке лингвистики и других гуманитарных наук, например, правоведения [Stubbs, Davis, 1921]; в стремлении математически смоделировать какой-либо статистически предсказуемый исторический процесс, такой, как деградация словаря или расхождение языка на диалекты; в обнаружении изоглосс [Campbell, 2013]; наконец, в анализе развития различных элементов концептосферы народа-носителя.

Обширное и детальное исследование предъявляет определенные требования к корпусной разметке; в частности, репрезентативность текста в рамках статистического анализа не в последнюю очередь обеспечивается лемматизацией корпуса, т.е. возможностью извлечения данных по словарным формам [Каримов, 2016]. Помимо лемматизации, для исторического корпуса важна экстралингвистическая разметка (метаданные), например, указание автора

и диалекта, различение года появления текста и года создания его манускрипта, включенного в корпус; разделение текстов корпуса по типам записи, в том числе с параллельным представлением текста в различных типах, как это сделано в [Medieval Nordic Text Archive (Menota)], где тексты представлены в манускриптовой, дипломатической и нормализованной формах письма. Не менее важны синтаксические пометы, в том числе указание части речи, или PoS-tagging. Такая разметка может принимать разные формы и является одним из самых популярных и востребованных на сегодняшний день тем научной работы в отрасли прикладной лингвистики [Silva, Silva, Rodrigues, 2015].

Невзирая на то, что на сегодняшний день имеется достаточное разнообразие исторических корпусов, в том числе размеченных, включая Хельсинкский корпус, Лингвистический атлас раннего среднеанглийского языка, Menota, такие корпусы включают далеко не весь спектр сохранившихся текстовых памятников, отражающих разнообразие того или иного языка в диахроническом аспекте. В этой связи у исследователя может возникнуть потребность в создании собственного корпуса. Хотя в настоящее время существует обширный инструментарий собственноручной разработки и аннотирования корпуса, включая General Architecture for Text Engineering, Atomic и др., подобный инструментарий либо предлагает полностью ручное внесение разметки, как Atomic, либо требует от пользователя значительных навыков программирования, как GATE. В целях автоматизации разметки, например, частеречной или лемматизации, можно воспользоваться существующими предварительно обученными моделями, такими, как TreeTagger, TNT, или разработанный в Копенгагенском университете лемматизатор на основе конечных автоматов [Karttunen, 2001]; однако для подобных алгоритмов отсутствуют модели, обученные на материале исторического языка. Необходимость автоматизировать процесс обусловлена тем, что полностью ручное размечивание корпуса является трудоемким, длительным и дорогостоящим процессом; так, создание сравнительного небольшого по объему Лингвистического атласа раннего среднеанглийского

языка потребовало финансирования и привлечения научных работников со стороны пяти учреждений: Научного совета Великобритании по гуманитарным наукам, Британской Академии, Фонда Леверхальма, Эдинбургского и Кейптаунского университетов [Laing, 2013].

Именно в этом ключе формулируется основная прикладная задача настоящего исследования: изыскать способ автоматизации частеречной разметки диахронического корпуса путем машинного обучения на основе предварительного пользовательского ввода сравнительно малого объема. Подобная формулировка ставит проводимую работу в контекст компьютерной и корпусной лингвистики как подразделов прикладной, в связи с чем считаем уместным начать обсуждение вопроса с рассмотрения специфики развития данных субдисциплин.

1.1. Особенности развития компьютерной и корпусной лингвистики в контексте сравнительно-исторического языкознания

Исторически подходы к обработке речи языкового материала значительно разнились в таких дисциплинах, как информационные технологии, электротехника, лингвистика и психология/когнитивная наука. Из-за таких расхождений данным вопросом занимается несколько разных, но во многом пересекающихся дисциплин: компьютерная лингвистика, обработка естественных языков как раздел информационных технологий, распознавание речи как проблема электротехники, а также компьютерная психолингвистика как одна из субдисциплин психологии. В данном пункте обобщается несколько исторических процессов, давших начало области изучения обработки речи и языка, а также компьютерной лингвистике в целом.

История данной конкретной области начинается с плодородного в плане интеллектуальных достижений и научного прогресса периода, наступившего после Второй мировой войны и давшего начало самому понятию компьютера. В этот период с 40-х до конца 50-х гг. велась интенсивная работа над двумя

основополагающими парадигмами: концептом автомата и вероятностными, или информационно-теоретическими моделями.

Автомат появился в 50-е годы, будучи доработанным вариантом предложенной А. Тьюрингом в 1936 году модели алгоритмического вычисления, которую сейчас многие считают фундаментом современной компьютерной науки. Благодаря работам Тьюринга У. Маккаллох и У. Питтс разработали свой нейрон [McCulloch, Pitts, 1943] – упрощенную модель нейрона как вычислительного элемента, описываемого средствами пропозициональной логики; затем на основе доводов Тьюринга С. Клини исследовал конечные автоматы и регулярные выражения [Kleene, 2016]. К. Шэннон [Shannon, 1948] применил вероятностные модели дискретных марковских процессов к языковым автоматам. Основываясь на идеях конечных марковских процессов, описанных в работе К. Шэннона, Н. Хомски в 1956 году впервые рассмотрел конечные компьютеры как способ описания грамматики, а также определил конечный язык как язык, порождаемый конечной грамматикой [Chomsky, 1956]. Эти первые модели в конечном счете стали основой теории формальных языков, которая использует инструментарий алгебры и теории множеств, определяя формальный язык как последовательность символов. К ней относится контекстно-свободная грамматика, впервые определенная Н. Хомски в 1956 году для естественных языков, но также независимо от него открытая Дж. Бакусом в 1959 году [Backus, 1959], а также его коллегами в 1960 году [Backus и др., 1962]; последний пользовался данным инструментом для описания языка программирования ALGOL.

Вторым основополагающим открытием этого периода стало развитие вероятностных алгоритмов обработки языка и речи, базисом которых стало другое исследование К. Шэннона: использование шумовых каналов и шифровальных методов для передачи языка такими средствами, как коммуникационные каналы и речевая акустика. К. Шэннон также позаимствовал у термодинамики понятие энтропии как способа измерить информационную

емкость канала, или информационное содержание языка, а затем первым в мире замерил энтропию английского языка, пользуясь вероятностными методиками.

Кроме того, именно в это время был разработан звуковой спектрограф [Koenig, Dunn, Lacy, 1946]; в области инструментальной фонетики были проведены фундаментальные исследования, заложившие основу последующих работ в распознавании речи.

В 1952 году исследователи лаборатории Bell Labs разработали статистическую систему, которая могла распознать любую из 10 цифр, произносимых одним и тем же говорящим [Davis, Biddulph, Balashek, 1952]. Так, уже в начале 50-х гг. появились первые устройства распознавания речи. В системе сохранялось 10 паттернов, формируемых индивидуально каждым говорящим и приблизительно кодирующих первые две гласные форманты в таких цифрах. Исследователи добились точности распознавания 97–99% за счет того, что система выбирала те паттерны, где наблюдался наивысший относительный коэффициент корреляции с вводными данными.

К концу 50-х и началу 60-х годов XX века исследования в области обработки речи и языка четко разделились на две парадигмы: символическую и стохастическую. Символическая парадигма зиждилась на двух направлениях исследований. Первое направление было представлено работами Н. Хомски и других исследователей теории формальных языков и порождающего синтаксиса, выпущенных в конце 50-х и до середины 60-х гг., а также в трудах многих лингвистов и специалистов по информационным технологиям, занимавшихся разработкой алгоритмов анализа, изначально реализовавшихся сверху вниз или снизу вверх, а затем также с применением средств динамического программирования. Одной из первых завершенных систем синтаксического анализа был Проект анализа трансформаций и дискурса (TDAP) Зелига Харриса, реализованный с июня 1958 по июль 1959 гг. в университете Пенсильвании [Longacre, Harris, 2006].

Вторым направлением стала новая область искусственного интеллекта. Летом 1956 года Джон Маккарти, Марвин Мински, Клод Шэннон и Натаниэль

Рочестер собрали группы исследователей в рамках двухмесячного семинара по вопросам нового концепта, который они решили назвать искусственным интеллектом (ИИ). Хотя в данной области всегда работало некоторое число исследователей, фокусирующихся на стохастических и статистических алгоритмах (включая вероятностные модели и нейронные сети), основное внимание уделялось логике и мышлению. Типичным примером такой работы является теоретик логики и универсальный решатель задач — продукты, разработанные и исследованные А. Ньюэллом и Г. Саймоном [Jurafsky, Martin, 2009].

К этому моменту уже были созданы первые системы понимания естественных языков. Эти простые системы работали в одном домене, сочетая подбор паттернов и поиск по ключевым словам с использованием простейшего эвристического анализа в целях разбора и формулировки ответов на вопросы. К концу 60-х были разработаны более формальные системы логики. Стохастическую парадигму в основном развивали кафедры статистики и электротехники. К концу 50-х гг. метод Байеса начали применять в решении задачи оптического распознавания символов. В. Бледсоу [Bledsoe, Browning, 1959] разработал систему распознавания текста на основе байесовских методов, где использовался большой словарь и производился расчет правдоподобия каждой наблюдаемой последовательности символов с учетом каждого слова в словаре; основным принципом было умножение вероятностей возникновения каждой буквы. Ф. Мостеллер и Д. Уоллес [Mosteller, Wallace, 1963] применили байесовы методы к решению задачи определения авторства статей в «Записках федералиста».

В 60-х гг. также появились первые серьезные и верифицируемые психологические модели обработки человеческого языка, основанные на трансформационной грамматике, а также первый онлайн-корпус: Brown Corpus, набор текстовых фрагментов общим объемом 1 миллион слов, взятых из 500 письменных текстов различных жанров (газет, романов, нехудожественной и академической литературы и т.д.). Этот корпус был создан в университете

Брауна в 1963–1964 гг. [Kučera, Nelson, 1967]; Уильям С. Й. Ван в 1967 году опубликовал свой DOC (Dictionary on Computer), онлайн-словарь диалектов китайского языка.

В следующий период наблюдался бум в исследованиях проблем обработки языка и речи, было разработано несколько исследовательских парадигм, которые до сих пор являются основными в данной области.

Стохастическая парадигма сыграла огромную роль в разработке алгоритмов распознавания речи, особенно в применении скрытых марковских моделей (СММ) и теорем каналов с шумами и декодированием, разработанных независимо Ф. Джелинеком, Л. Балем, Р. Мерсером и др. Исследовательского центра IBM им. Томаса Дж. Уотсона, а также Бейкером в Университете Карнеги Меллон; последний вдохновлялся работой Баума и др. в Институте оборонного анализа, Принстон. Bell Laboratories в составе AT&T также стала одним из ключевых центров исследования распознавания и синтеза речи; в книге название книги [Rabiner, Juang, 1993] подробно описывается широкий спектр таких работ.

Логической парадигме дала начало работа А. Колмерауэра и коллег в области Q-систем и метаморфозных грамматик [Colmerauer, 2006]. Эти работы предшествовали созданию языка программирования Prolog, а также грамматик определенных придаточных [Pereira, Warren, 1980]. Независимо от них работа М. Кея [Kay, 1979] по вопросам функциональной грамматики, а также вышедшая вскоре после нее работа Дж. Бреснана и Р. Каплана [Bresnan, Kaplan, 1982], посвященная лексической функциональной грамматике (ЛФГ), продемонстрировали важность унификации структур элементов.

Область вопросов понимания естественного языка также зародилась в течение этого периода, а начало было положено системой SHRDLU, разработанной Т. Виноградом. Данная система симулировала робота в мире, созданном из игрушечных блоков [Winograd, 1972]. Программа могла понимать текстовые команды на естественном языке («Поставь красный кирпич на маленький зеленый») такой сложности, которая до этого казалась немыслимой. Его система также стала первой попыткой построить расширенную и

продвинутую (для того времени) грамматику английского языка на основе системной грамматики М. Халлидея. Модель Т. Винограда дала понять, что достигнутого понимания проблем синтаксического анализа было достаточно, чтобы сосредоточиться на семантике и дискурсе.

Роджер Шэнк, его коллеги и студенты в рамках того, что называют Йельской школой, разработали несколько программ понимания языка, где основное внимание уделялось концептуальному знанию – сценариям, планам и целям, а также организации человеческой памяти [Carbonell, Cullingford, Gershman, 1981; Lehnert, 1977; Schank, Riesbeck, Riesbeck, 2017]. В такой работе часто использовалась сетевая семантика [Norman и др., 1975; Quillian, 1968; Wilks, 1975], а также начали применять понятие роли, сформулированное Ч. Филлмором [Fillmore, 1968], в представлении такой семантики [Simmons, 1973].

Логическая парадигма и парадигма понимания естественных языков объединились в системах, где предикатная логика использовалась в качестве семантического представления, например, в LUNAR — системе, формулирующей ответы на вопросы [Woods, 1979].

Парадигма моделирования дискурса основное внимание уделяла четырем ключевым областям дискурса. Б. Грос и ее коллеги начали изучать субструктуры и фокус дискурса [Grosz, 1977]; несколько исследователей стали работать над разрешением вопросов формирования автоматических ссылок [Hobbs, 1978]; была разработана модель убеждений, желаний и намерений, на основе которой велось исследование речевых актов по принципам формальной логики [Perrault, 1980].

В следующем десятилетии в лингвистику вернулись два класса моделей, утративших популярность в конце 50-х и начале 60-х гг., в основном из-за теоретических доводов против таких моделей, например, рецензией Н. Хомски на работу Б. Скиннера «Вербальное поведение» [Chomsky, 1959]. Первым таким классом являются модели с конечным числом состояний: после публикации работ, посвященных конечной фонетике и морфологии [Kaplan, Kay, 1981] и

конечным моделям синтаксиса [Church, 1980], таким моделям снова стали уделять внимание.

Вторым трендом этого периода стало то, что называют «возвращением эмпиризма»; здесь наиболее примечателен рост популярности вероятностных моделей в решениях задач обработки речи и языка, где большим влиянием пользовались исследования вероятностных моделей распознавания речи, проводимые Исследовательским центром IBM им. Томаса Дж. Уотсона.

Вероятностные методы и другие подходы на основе данных стали также применяться в решениях таких проблем, как разметка частей речи, анализ синтаксиса и вопросы омонимии, семантика. Это эмпирическое направление сопровождалось также акцентом на оценке моделей, которая основывалась на применении данных о накопленном опыте, разработке квантитативных оценочных метрик, а также на сравнении эффективности по таким метрикам с показателями ранее опубликованных исследований.

Кроме того, начались масштабные работы по исследованию проблем порождения естественных языков.

К концу XX века стали очевидны значительные изменения в данной области. Во-первых, вероятностные модели и модели на основе данных стали стандартом всех исследований обработки естественного языка. Алгоритмы синтаксического анализа, разметки частей речи, разрешения ссылок и обработки дискурса стали включать вероятностные компоненты, а также оперировать методологиями оценки, позаимствованными из областей распознавания речи и информационного поиска. Кроме того, компьютеры стали работать быстрее, увеличилась их память, что дало возможность применять некоторые подобласти обработки речи и языка коммерчески, в частности, появились домашние программы распознавания речи, проверки орфографии и грамматики. Алгоритмы обработки речи и языка стали применяться в такой области, как аугментативная и альтернативная коммуникация (ААК). Наконец, популяризация и развитие Интернета обусловили необходимость поиска и извлечения информации на лингвистической основе.

Эмпирический подход, зародившийся в конце 90-х, стал бурно развиваться в новом столетии. Это ускорение было в основном обусловлено тремя синергетическими тенденциями.

Во-первых, под эгидой Консорциума лингвистических данных (КЛД) и схожих организаций был обеспечен доступ к большим объемам устного и письменного материала. Важно отметить, что к таким материалам относятся аннотированные корпуса текстов: Penn Treebank [Taylor, Marcus, Santorini, 2003], Prague Dependency Treebank, PropBank [Palmer, Gildea, Kingsbury, 2005], Penn Discourse Treebank [Miltsakaki и др., 2004], RSTBank [Carlson и др., 2001] и TimeBank [Pustejovsky и др., 2003]; в этих базах обычные источники текста дополнены множественными слоями синтаксической, семантической и прагматической разметки. Доступность таких ресурсов стимулировала решение сложных традиционных проблем, таких, как синтаксический или семантический анализ с помощью контролируемого машинного обучения. Эти ресурсы также способствовали появлению дополнительных методов оценки синтаксического анализа [Sang, Dejean, 2001], информационного поиска, снятия омонимии [Kilgarriff, Palmer, 2000], генераторов ответов на вопросы [Voorhees, Tice, 2001], а также реферирования [Dang, 2006]

Во-вторых, уделяя большее внимание обучению машин, лингвисты стали более серьезно взаимодействовать со специалистами по статистическому машинному обучению. Такие методы, как опорные векторы [Boser, Guyon, Vapnik, 1992; Vapnik, 2000], максимальная энтропия и ее эквиваленты в виде многочленных логистических регрессий [Berger, Pietra Della, Pietra Della, 1996], а также графические байесовы модели [Pearl, Judea, 1988] стали обычной практикой компьютерной лингвистики.

В-третьих, широкое распространение высокопроизводительных вычислительных систем способствовало подготовке и развертыванию систем, которые никто не мог представить себе десять лет назад.

Наконец, практически в конце этого периода начали уделять повышенное внимание статистическим подходам, при которых человеческий контроль

практически отсутствует. Прогресс в статистических подходах к машинному переводу [Brown и др., 1990] и тематическому моделированию [Blei, Ng, Edu, 2003] показал, что эффективные приложения можно конструировать на основе систем, обученных только с помощью неразмеченных данных. Кроме того, затраты и сложности, связанные с созданием качественно аннотированных корпусов, стали фактором, ограничивающим контролируемый подход ко многим вопросам.

Методы обучения без учителя (англ. unsupervised learning), скорее всего, будут становиться все более и более популярными.

Несмотря на то, что компьютерная лингвистика как отрасль науки развивается уже более полувека, одна из ее особенностей остается неизменной: в большинстве своем перечисленные работы по компьютерной лингвистике в той или иной мере полагаются и/или ссылаются на психологические исследования проблематики человеческого мышления, причем в последние годы данная тенденция усиливается. Конечно, понимание того, как человеческий мозг обрабатывает языковые данные, уже само по себе является важной научной задачей и относится к когнитивной науке в целом. Однако понимание данных процессов зачастую может быть полезно при совершенствовании машинных моделей языка.

Использование природных моделей особенно эффективно в решении чисто гуманитарных задач. Так, цель систем распознавания речи заключается в том, что выполнять задачу, с которой регулярно сталкиваются секретари судов: транскрибировать устный диалог. Так как люди уже хорошо справляются с этой задачей, имеет смысл обращаться к уже готовому природному решению. Более того, так как важнейшей областью применения систем обработки речи является взаимодействие человека и компьютера, имеет смысл скопировать решение, которое будет работать в привычной человеку манере.

Именно обращение к когнитивной специфике языковой деятельности, к алгоритмам обработки естественного языка в рамках умственно-речевой деятельности его пользователей может стать потенциальным решением задачи

автоматизации частеречной разметки в условиях ограниченности материала – задачи, в отношении исторического корпуса еще не решенной окончательно, о чем говорится в следующем пункте.

1.2. Обзор текущей ситуации в решении задач частеречной разметки исторического корпуса

Проблематика частеречной разметки не является новой: автоматизация разметки корпуса применялась еще в начале 80-х гг. при построении Пеннсильванского корпуса английского языка. Обучение алгоритмов, применимых к английскому языку, преимущественно основано на использовании современного языкового материала, доступного в большом объеме; так, модель Леона Дерчински и др. [Derczynski и др., 2013], разработанная командой Шеффилдского университета, обучена на подборке сообщений из социальной сети Twitter; теггер Стэнфордского университета использует базу данных Penn Treebank; а в отношении русского языка имеется морфологический анализатор с функцией стеммирования/лемматизации и частеречной разметки, созданный командой Яндекса [Большакова и др., 2017]. Однако данные модели не всегда успешно применимы к историческим языкам по различным причинам, которые будут рассмотрены далее в п. 1.3 «Среднеанглийский и древнескандинавский языки: особенности частеречной структуры и орфографии», ввиду чего при работе с историческим корпусом возникает необходимость в новых алгоритмах. В этом пункте рассматриваются последние разработки в данной области.

В связи с тем, что оба языка – среднеанглийский и древнескандинавский, – о которых пойдет речь, относятся к германской группе индоевропейской языковой семьи, имеет смысл в первую очередь обратиться к имеющимся наработкам, применимым именно к германским языкам. Так, М. Колева и др. [Koleva и др., 2017] применили два различных алгоритма машинного обучения – обучение по памяти (англ. memory-based learning) и условно-случайное поле

(англ. conditional random field) – к предварительно размеченному корпусу средневекового верхненемецкого языка; утверждая, что оба алгоритма, относясь к категории так называемого «ленивого обучения» (англ. lazy learning), т.е. не выполняя никакой генерализации по используемой в обучении базе данных, обеспечивают большую точность и полноту по сравнению со скрытыми марковскими моделями, деревьями принятия решений и т.д., авторы достигли среднего показателя точности междиалектной классификации (т.е. с обучением на одном диалекте и проверкой на другом) в нескольких экспериментах, отличающихся постановкой, на уровне 88%; при межжанровой классификации достигнута точность в 77%; оба результата можно назвать значительным прогрессом по сравнению с показателем 75%, полученным ранее, в 2016 году, с помощью классификатора RFTagger [Barteld, Schröder, Zinsmeister, 2015]. Нужно отметить, что используемые авторами классификаторы обучались на морфологических N-граммах, но не на полном символьном составе слова.

Я. Байз и Я. Бота, Оксфордский университет и Google Inc., соответственно, предложили модель обучения без учителя, применимую для частеречной разметки языков с ограниченным доступным материалом [Buys, Botha, 2016], что в принципе схоже с основной сложностью рассматриваемой в настоящем диссертационном исследовании задачи, так как исторический корпус также отличается малыми размерами (до 100 тысяч слов в отдельно взятый период развития среднеанглийского языка). Их модель основывалась на проецировании данных уже размеченного текста на параллельный ему неразмеченный текст на другом языке, позволяя существенно ограничить подмножество применимых в том или ином предложении частеречных тегов. В то время как достигнутые комбинированием скрытой марковской модели и модели Всаби результаты впечатляют (точность до 80% при использовании размеченного английского текста в качестве основного), данный подход нельзя считать подходящим в рамках настоящего исследования, так как его цель заключается в обучении алгоритмов на части языкового материала, подлежащего разметке, без применения параллельных текстов. Тем не менее, морфологическая

комплексность рассмотренных ими 11 языков, в том числе финского и болгарского, позволяет частично обратиться к описанной данными исследователями теории.

Ахим Штайн (Штуттгартский университет) описал проблематику аннотирования текстов на старофранцузском с применением амстердамского корпуса, обращаясь при этом к алгоритму TreeTagger, основанному на деревьях принятия решений; достигнута невероятно высокая точность определения части речи: около 95% [Stein, 2008].

Кристина Марко (Университетский колледж Йовик [Marco, 2012]), воспользовавшись инструментом лингвистического анализа Freeling [Padró, 2011] с открытым исходным кодом, разработала модель частеречной разметки текста на современном норвежском языке, обеспечивающую точность до 97,1% при определении первичных характеристик (т.е. самой части речи) и до 92% при определении вторичных и третичных свойств (например, рода, числа или падежа). В то время как ее результаты значимы как в рамках настоящей работы, так и с точки зрения прикладной лингвистики в целом ввиду малой исследованности современных скандинавских языков методами компьютерной филологии, их нельзя назвать напрямую применимыми к текущему исследованию в связи с той значительной разницей, которую можно наблюдать между современным и старым норвежским (см. п. 2.5).

Клаудио Якобини и др. [Iacobini C., Rosa A. De, Schirato G., 2014] предложили стратегию частеречного аннотирования диахронического корпуса итальянского языка MIDIA, которая также основана на вероятностном моделировании с помощью TreeTagger – алгоритме, работа которого основана на вероятности принадлежности токена к той или иной части речи в зависимости от частеречного окружения; недостатком такого подхода является неспособность алгоритма справляться с классификацией слова вне его контекста, так как TreeTagger не задействует внутренние особенности самого слова, т.е. его морфологию. Таким образом, даже весьма низкий уровень ошибок анализатора

(3,72%), полученный на материале MIDIA, нельзя считать релевантным в рамках настоящего исследования.

Интересна работа Рафаэля Гуаращи, создавшего анализатор латинских текстов на основе материала онлайн-словаря Wiktionary; работа его алгоритма основана на применении нейросетевой модели усредненного перцептрона (англ. averaged perceptron), при этом достигнута точность частеречной классификации порядка 80% при использовании обучающей выборки размером всего в 10000 слов, что однозначно является хорошим результатом ввиду высокой морфологической комплексности латыни [Guarasci, 2017]; а так как материал для обучения не являлся контекстуализированным (использовался словарь, а не текст, т.е. извлечение N-грам из токенов не представлялось возможным), вкупе с фактом отсутствия нормализации орфографии материала, задействованный подход, который дополнялся методикой посимвольного сопоставления, может быть полезен в настоящем исследовании.

Наряду с латынью изучалась и проблематика частеречного аннотирования древнегреческого языка. Дж. Челано и др. [Celano, Crane, Majidi, 2016] сравнили эффективность пяти различных алгоритмов аннотирования: Mate Tagger, основанный на переходных моделях; TNT, основанный на марковских моделях второго порядка; RFTagger, основанный на скрытых марковских моделях; OpenNLP, работающий по принципу максимальной энтропии; и NLTK Unigram, который присваивает примеру из проверочной выборки тот класс, с которым аналогичный пример чаще всего ассоциировался в обучающей выборке. Наибольшей точностью отличается первый алгоритм: 88% для древнегреческого языка при 5-проходной перекрестной проверке. Следует отметить, что авторы не вносиили никаких изменений ни в сами алгоритмы, ни в репрезентацию текста: их цель заключалась в сравнении эффективности алгоритмов, в связи с чем данная работа не является однозначно релевантной в настоящем исследовании.

Возвращаясь, наконец, к английскому языку, нельзя не отметить работу Муна и Болдриджа [Moon, Baldridge, 2007] – разработку алгоритма частеречного аннотирования без учителя, который основан на сопоставлении и проецировании

двух параллельных текстов на английском языке, отличающихся диахронически; таким образом, для разметки текста на среднеанглийском языке задействуется параллельный ему текст на современном английском, уже содержащий разметку (в качестве основного материала использовались исторически разнесенные переводы Библии); сопоставление и коррекция выполняются с помощью вероятностной модели, похожей на теорему Байеса, в то время как для репрезентации данных используются биграммы (т.е. работа разметчика зиждется на расчете вероятности события A при условии возникновения события B). Высокая точность на уровне 95% говорит об эффективности данного подхода, однако малая доступность параллельного диахронического текста, сложность его подготовки и необходимость доработки алгоритма под особенности каждого конкретного исторического языка вкупе с контекстуальностью вероятностных моделей на биграммах ставит под сомнение применимость этой концепции в текущей работе.

И. Ян и Ж. Айзенштайн из Джорджийского технологического института (Атланта, США) разработали новаторский алгоритм частеречной разметки корпуса на историческом английском языке, конкретнее – раннем новоанглийском языке – без нормализации орфографии [Yang, Eisenstein, 2016]. Их подход основан на так называемом принципе доменной адаптации, который позволяет использовать статистические данные со-возникновения (англ. co-occurrence) признаков для преображения целевого материала в такой вид, который будет ближе к материалу обучения (т.е., например, использовать современный язык для обучения анализатора работе с историческим языком). Для этого используется метод структурных соответствий: искусственная генерация большого числа бинарных задач, нацеленных на установление так называемых «ключевых» признаков (англ. pivot features); статистические данные со-возникновения таких признаков позволяют ассоциировать ключевые признаки в разных доменах (например, в современном английском и ранненовоанглийском языках). Далее задействуется подход, названный авторами «встраиванием признаков» (англ. feature embedding), позволяющий

добиться высокой плотности и низкой размерности векторного представления каждого примера. В качестве базовых алгоритмов обучения используются методы опорных векторов (со специально подобранными параметрами регуляризации) и двухнаправленные марковские модели с максимальной энтропией; пользуясь признаковыми шаблонами А. Ратнапархи [Ratnaparkhi, 1996] и встраиванием признаков в целях снижения вероятности ошибок авторам удается добиться высокой точности классификации на уровне 97% при диахроническом разбросе учебного и проверочного материала в более чем 150 лет. Такой показатель делает данную работу основным опорным материалом при создании нашего алгоритма машинного обучения в связи с необходимостью обработки корпуса с большим диахроническим охватом, а использование тех же методов (опорные векторы и СММ) повышает ценность труда И. Яна и Ж. Айзенштайна как источника идей для настоящего диссертационного исследования. Тем не менее, стоит отметить, что авторам был доступен значительный объем материала для обучения, а сам алгоритм проверялся лишь на ранненованглийском языке (16 век и позднее), что ограничивает применимость метода в настоящем исследовании.

Если же говорить о древнескандинавском языке, то нельзя не отметить работу Е. Рёгнвальдсона и С. Хельгатоухтир [Loftsson, Helgadóttir, Rögnvaldsson, 2011], в которой удалось применить аннотатор TNT, обученный на материале современного исландского языка, к древнескандинавскому, добившись без внесения каких-либо корректировок точности в 90,36%, при этом не выполнялось никакой параллелизации и никакого проецирования данных; впрочем, близость исландского к его языку-предку отмечалась еще Расмусом Раском [Rask, 1843]. Корректировка достаточно большого предварительно размеченного отрезка корпуса на древнескандинавском языке и последующее повторное обучение аннотатора на нем же позволяет значительно повысить точность до 92,65%. Прочие алгоритмы обучения авторами не оценивались, попыток комбинирования с другими не производилось. В то время как полученные результаты интересны, а теоретическое обоснование и обсуждение

стоит признать полным и обширным, данная работа не имеет существенного значения для настоящего диссертационного исследования ввиду расхождения как задач, так и изначальных условий и постановки проблематики. Прочих значимых исследований в области аннотирования текстов на древнескандинавском языке автору диссертации обнаружить не удалось.

Отметим, что текущее состояние вопроса обусловливает уникальное положение настоящего исследования и его научную новизну в отечественной компьютерной и корпусной лингвистике при изучении исторических текстов. В то время как вопросы применения машинного обучения к корпусному материалу российскими учеными рассмотрены достаточно широко, в основном они затрагивают проблематику классификации текстов и информационного поиска, практически не касаясь задач частеречной разметки. Из публикаций последних пяти лет в области корпусной лингвистики и МО интерес представляют работы А.О. Шелманова и М.К. Каменской (ИСА ФИЦ ИУ РАН), где исследованы подходы к определению ролевых структур высказываний, использующих принципы машинного обучения с частичным привлечением учителя [Шелманов, Каменская 2017]. Г.Г. Новиков и И. М. Ядыкин (НИЯУ МИФИ) запатентовали программу для ЭВМ, обеспечивающую нормализацию корпуса текста в целях последующего дистрибутивного анализа; программа использует алгоритм Doc2Vec, реализованный на языке программирования Python, и предназначена для предварительной бинаризации корпуса текстов на естественном языке для последующего нечеткого поиска текстов по подобию [Новиков, Ядыкин, 2016]. Классификационную задачу рассматривает М.И. Попков (МГУ им. Ломоносова) в рамках более широкой проблемы формирования базы знаний предприятия [Попков, 2014]; к классификации текста с помощью приводимых далее в настоящем исследовании моделей на основе опорных векторов обращается Д.А. Гапонов (АлГУ) [Гапонов, 2017]. Д.В. Климов (МТИ) предлагает авторский метод уменьшения признакового пространства матрицы объектов в задаче классификации текстов с применением различных алгоритмов машинного обучения, в частности, логистической регрессии [Климов, 2017], которая

задействуется в том числе в представляемом исследовании. В то время как данные работы сами по себе значимы в общем контексте специальности 10.02.21, с одной стороны, и в существенной мере перекликаются с настоящей работой с точки зрения общности рассматриваемой проблематики, в частности, уделяя внимание таким проблемам, как обеспечение векторного кодирования малой размерности и применение классификационных алгоритмов обучения с учителем на материале корпуса относительно небольшого объема, основной их целью является классификация документов (текстов) и информационный поиск, а не отдельных неделимых элементов текста (лексем). По схожей причине в рамках данной работы неприменимы результаты исследования В.В. Артюхина (ФГБУ ВНИИ ГОЧС (ФЦ), изучавшего возможности применения методов машинного обучения при работе с литературными источниками, в частности, приложения алгоритма Луна: его анализ касается проблематики кластеризации, а не классификации [Артюхин, 2015].

Несомненный интерес для корпусной лингвистики представляет исследование А.Г. Асташкина и К.В. Чувилина (ИФТИ и МФТИ)), исследовавших проблематику частичной формализации текстового документа в целях описания формальной грамматики; в их работе [Асташкин, Чувилин, 2017] представлен обученный на корпусе научных статей в формате LATEX алгоритм, автоматизирующий внешнее описание элементов синтаксиса. В то время как синтаксическое описание текста само по себе содержит указание частей речи (точнее, членов предложения), оно служит иным задачам, нежели классификация по частеречному признаку, выполняется по иным принципам, не обеспечивает человеко-читаемости (англ. human readability) выходных данных и не соответствует целям настоящего исследования.

С перспективными целями данной работы отчасти пересекается проект, представленный в статье С. А. и Е.В. Полицыных [Полицын, Полицына, 2019], где описывается реализация комплекса инструментов управления корпусами текстов при решении задач компьютерной лингвистики, в частности, создания, модификации и признакового выделения субкорпусов из крупного корпуса

текстов (т.е. задач, схожих с выполняемыми платформой SketchEngine), что, с одной стороны, позволит обратиться к цитируемой работе в рамках будущей разработки корпусного редактора (цель заявлена в [Каримов, Акинин, 2017]), с другой – неприменимо напрямую на текущей стадии исследования.

Непосредственный интерес с точки зрения настоящей работы представляют три опубликованные недавно и проиндексированные в Российском индексе научного цитирования (РИНЦ) статьи. В первую очередь отметим публикацию А. Молдагуловой и Г. Бектемысовой [Молдагулова, Бектемысова, 2017], в которой рассказывается о применении скрытых марковских моделей в целях автоматизации частеречной классификации лексем в корпусе казахского языка. В то время как основная поставленная задача в этом и нашем исследованиях совпадает, непосредственный перенос результатов и методов не представляется возможным в связи с а) использованием отличного языкового материала; б) применением СММ, от которых в рамках настоящей диссертационной работы решено отказаться.

С позиции непосредственной переносимости результатов и методик более значимой представляется статья И. А. Молошникова и др. о применении глубоких нейронных сетей в целях создания морфологической разметки на материале корпуса русского языка СинТагРус при точности порядка 98% [Молошников и др., 2017]; по мнению авторов, «полученные результаты являются лучшими среди рассматриваемых подходов машинного обучения для русского языка при использовании только корпусной информации без дополнительных словарей и правил» [Там же]. В работе используется как пословное, так и посимвольное представление предложений, причем последнее интересно в том числе потому, что посимвольной векторизацией решается задача репрезентации текста на вход алгоритмам машинного обучения в настоящей работе, см. п. 1.3.1 «Методы векторизации текста, применимые к историческому тексту». Применимость описанного метода в настоящей работе ограничивается тем, что оба представления задействованы единовременно, в то

время как методология настоящего исследования ограничивается одним подходом к векторизации.

Корпус СинТагРус используется в еще одной работе, посвященной непосредственно автоматизации частеречной разметки и частично перекликающейся с настоящим диссертационным исследованием методологически, в частности, за счет обращения к методу опорных векторов. Речь идет о статье Д.В. Тарасова (ПГУ) и Н.А. Романова (МГУ), посвященной вопросам морфологической разметки и определения частей речи во флективных языках [Тарасов, Романов, 2017]. Используя собственную реализацию применяемых алгоритмов на языке C/C++, авторы достигли точности частеречной классификации на уровне 87–95%, при этом их конечной задачей является разработка программного средства распознавания текста на русском языке, т.е. непосредственно PoS-разметка преподносится как промежуточный этап. Опираясь преимущественно на опорные векторы, исследователи анализируют линейную разделимость или ее отсутствие как одну из фундаментальных характеристик качества имеющегося набора данных, описывают различные методики нормализации выборки в целях обеспечения такой разделимости, рассматривают задачу присвоения частеречных помет как комплекс бинарных подзадач и используют ограниченный набор собственно классов; все это сближает их работу с настоящим исследованием, в особенности применение малого набора классовых помет, см. п. 2.1 «Корпусный материал и его подготовка». При этом о непосредственной переносимости их методики речи не идет: векторизация, задействованная в их работе, опирается на выделение приставок и окончаний, что, на наш взгляд, не вполне применимо к материалу древнегерманских языков в связи с их сильной внутрикорневой флексией, а к материалу среднеанглийского языка – еще и по причине значительной орфографической вариации; предварительное выделение регулярных морфов, например, с помощью алгоритма Morfessor, также не представляется возможным.

Таким образом, поставленная в рамках настоящего исследования задача автоматизации разметки корпуса на среднеанглийском и древнескандинавском языках не решена в полной мере зарубежными исследователями и не затронута напрямую в отечественной науке; при этом нельзя не отметить ценность имеющегося российского и международного научного опыта в разработке схожей проблематики и необходимость обращения к нему.

1.3. Среднеанглийский и древнескандинавский языки: особенности частеречной структуры и орфографии

Рассмотрение вопросов частеречной разметки текста корпуса невозможно без тщательного теоретического и практического анализа специфики грамматического оформления частей речи и связанных с ними орфографических особенностей, так как именно письменная реализация различий между именами и глаголами, самостоятельными и служебными словами, основными и вспомогательными членами предложения определяет функциональную специфику, методологию, процедуру и алгоритмы автоматической разметки. Решение подобной задачи в первую очередь требует адекватной входному тексту репрезентации в машиночитаемом виде, обрабатываемом выбранными алгоритмами и/или их ансамблем. С одной стороны, такая репрезентация должна обеспечивать дифференцирование частей речи уже на входе по тем или иным признакам, с другой – не приводить к возникновению неоднозначных ситуаций, когда отраженные во входных данных признаки могут одновременно охватывать различные части речи. В этой связи следует подробнее изучить характерные особенности самой частеречной структуры исследуемых языков и их орфографическую реализацию. Отметим, что последняя в значительной степени зависит от фонетических изменений в языке.

В рамках анализа имеет смысл рассмотреть общее положение двух сравниваемых языков в генеалогической структуре. Речь идет о германских языках – группе в составе индоевропейской языковой семьи, развившейся из

протогерманского языка, который в свою очередь отделился диалектально от протоиндоевропейского в районах к северу и западу от р. Висла ок. 1000 г. до н.э. [König, Auwera, 2002, p. 2] Германский вокабуляр содержит небольшое количество общеиндоевропейских когнатов, в основном в списке Сводеша [Larsson, 2013, p. 5]. Ключевым отличием германских языков от прочих членов индоевропейский семьи является фонологическое изменение, обусловившее переход части взрывных во фрикативы с их частичным последующим озвончением; принцип, описывающий такое изменение, иногда называют законом Гримма-Вернера [Расторгуева, 2003, с. 28].

Сама германская группа делится на три подгруппы: западную, куда помимо верхне- и нижненемецких, а также франконских говоров, вошли англо-саксонские диалекты племен, переселившихся с полуострова Ютландия на Британские острова в середине V в. н.э., впоследствии сформировавшие то, что сейчас называют древнеанглийским языком [Иванова, Чахоян, Беляева, 2006, с. 10]; вымершую к настоящему времени восточную [Clackson, 2007] и северную подгруппу, развившуюся из протоскандинавского языка, который затем трансформировался в древнескандинавский за счет процессов умлаута и преломления [Haugen, 2013].

В настоящей работе исследуются среднеанглийский и древнескандинавский языки. Обсуждение и сравнение языков начнем со второго в связи с тем, что он по сравнению со среднеанглийским более консервативен и относится к несколько более раннему периоду; значительная часть фонетической, грамматической, орфографической специфики является общей для древнескандинавского и древнеанглийского языков, но утрачивается в среднеанглийском. Ввиду этого считаем наиболее рациональным анализировать среднеанглийский с позиции его отличий.

Само определение **древнескандинавского языка** остается несколько спорным. Общепризнано то, что данный язык представлял собой разнообразный диалектный континуум, покрывавший достаточно крупную территорию, включавшую юг и западное побережье Норвегии, Исландию (с конца IX в.) и юг

Гренландии, Фарерские и Шетландские острова, Ютландию, Швецию, простираясь к востоку вплоть до юго-запада современной России [Bandle, 2002]. В начале второго тысячелетия уже наметились, хотя, возможно, еще не осознавались носителями диалектные различия между группами западных и восточных говоров, а также своеобразного «изолята» — гутнийского языка, статус которого (язык или диалект) до сих пор является предметом споров, не в последнюю очередь из-за неполного проявления общих северогерманских фонетических процессов. Некоторые исследователи древнескандинавского языка таковым признают только норвежско-исландский язык того времени, утверждая, что шведский, датский и гутнийский восходят к восточно-северному германскому языку [Jahr, Lorentz, 1993]; другие данный, более единообразный по сравнению с восточными говорами, язык называют древне- или старонорвежским (отметим, что Исландия была колонизирована именно выходцами из Западной Норвегии (норв. Vestlandet) в конце эпохи викингов [Haegstad, Torp, 1909], и до сих пор из всех континентальных языков именно западные норвежские диалекты ближе всего к исландскому фонетически); в связи с этим возникает еще одна терминологическая коллизия: название «старонорвежский язык» (gammalnorsk) может употребляться также в отношении собственно норвежского, рассматриваемого уже как самостоятельный, отдельный от исландского, язык периода 1350–1500 гг. н.э.; однако тот же самый период обозначается также как «средненорвежский» (mellomnorsk) [Magnus, 1993]. Однозначно выделение трех ключевых диалектных континуумов: западного и восточного древнескандинавских, а также гутнийского. Позднее наиболее значимые изоглоссы проходили по границам Дании и Швеции (Скона), разделившихся по признаку третичного перебоя согласных, вызвавшего озвончение, которое сейчас наблюдается только в датском языке, ср. норв. *kjøp* и дат. *kjøb*; гутнийских владений в Швеции, Трёнделага в Норвегии, западного норвежского побережья. Датский переход был настолько значимым, что языки периода 1100—1300 гг. уже делят на северные и южные скандинавские, относя к последним только датский [McTurk, 2005]. В рамках настоящей работы

рассматривается древнескандинавский язык как совокупность всех скандинавских диалектных континуумов в период с вырождения протоскандинавского языка-основы до общепринятой точки распада на отдельные языки, т.е. с IX по XIII вв. н.э., за исключением текстов на гутнийском языке/диалекте.

Если говорить об этом языке, не вдаваясь в фонетико-морфологическую специфику его диалектов, то необходимо отметить следующие ключевые особенности.

От протоскандинавского языка, использовавшего, подобно древнеанглийскому языку, старший футарк и оставившего корпус из 127 рунных надписей, преимущественно бустрофедоном [Bandle, 2002], древнескандинавский унаследовал руническую запись, которая преобразилась в младший футарк; здесь же отмечаем ключевое отличие от среднеанглийского языка: хотя большая часть дошедших до нас письменных памятников использует уже латиницу, что обусловлено влиянием христианства на становление скандинавской государственности ближе к окончанию эпохи викингов, рунная запись использовалась в церемониальных целях вплоть до конца XIII века [Schulte, 2005]. Как и в случае с английским языком, руны повлияли на латинскую письменность: скандинавы сохранили торн и эт; в корпусной дипломатической и нормализованных записях также используются акценты «акут», огóнек и ʀ-символ, обозначавший особый велярный звук, появившийся еще в протоскандинавском языке в результате ротацизма протогерманского /z/ (впрочем, слившийся с обычным /r/ ближе к окончанию древнескандинавского периода во всех диалектах) [Стеблин-Каменский, 1953].

С точки зрения вокализма полностью симметричная пентавокалическая система (5 коротких гласных и 5 соответствующих им долгих плюс 3–4 дифтонга) скандинавского языка-основы перешла в несимметричную систему из 9 монофтонгов с неполными парами по долготе. Причиной этого явления стали умлауты трех типов (a-, i-, u-умлауты в порядке старшинства) – явление, схожее с i-умлаутом западногерманских языков, однако более широкое, вызываемое

большим числом вокалических финалей. Кроме того, многие слова подверглись синкопированию – сокращению долгих финалей и исчезновением кратких; ср. протосканд. **magurʀ*[1] и древнесканд. *mǫgr*, «родственник мужеска полу» – u-умлаут, протосканд. **hurna* и древнесканд. *horn* – a-умлаут и синкопа; протосканд. **gastiʀ* и древнесканд. *gestr*, «гость» – i-умлаут. Еще одним важным фонетическим процессом стало преломление – изменение звукообразовательного качества гласной фонемы *e* под влиянием концевых гласных *a* и *u*, например, **erþu – jǫrð* (Земля). Все эти фонетические изменения повлияли на орфографию языка, в том числе с точки зрения форморазличения; так, i- и u-умлаут до сих пор несут грамматическую функцию, например, норв. *mann/menn* («мужчина» в ед. и мн. числе); исл. *tala/tölum* («говорить» в инфинитиве и 2 лице мн.ч. наст. вр.) – явление, которое называют морфофонологическим умлаутом [Hagland, 1978].

Еще одним интересным вокалическим феноменом, повлиявшим на орфографию и письменное различение частей речи, можно назвать ликвидацию зияний (гиатуса); так, кластер /e:a/ перешел в /ja:/, кластер /e:u/ – в /jo:/ или /ja:/, например, **séa – sjá; *tréum – trjóm*. Наконец, при суффиксации отмечается выпадение присутствующих в начальных формах безударных гласных: **dróttin – dróttnar* (владыка).

Рассуждая о древнескандинавском консонантизме с точки зрения его влияния на орфографию и частеречное деление, необходимо отметить прогрессивную ассимиляцию согласных, в результате которой звук /r/ утратил свою слогообразовательную функцию, так как выпал из кластеров с s и сонорантами, приведя, впрочем, к удлинению последних в случаях, когда основа оканчивалась кратким, безударным словом: **laksr – laks; *stólr – stóll; *agnr – agn*, но **jǫtunr – jǫtunn*. Наблюдалась и регрессивная ассимиляция, в частности, слияние консонант /d, ð/ с последующим звуком /t/: **fǿdt – fǿtt, *kallaðt – kallat* [Там же].

[1] Здесь и далее звездочкой «*» помечаются реконструированные слова.

Наконец, нельзя не отметить выпадение аппроксимант в определенных условиях, в частности там, где под влияние i-умлаута попадает кластер *ja*, преобразующийся не в *je*, как следовало бы ожидать, а в *i*, см. пример со словом *fjǫrðr*, «фьорд», далее.

С точки зрения частеречной и морфограмматической структуры древнескандинавский мало отличается от прочих германских языков своего времени. Традиционно в древнескандинавской филологии принято говорить о десяти частях речи: существительных, прилагательных, местоимениях, артиклях, числительных, глаголах, наречиях, предлогах, союзах и междометиях. В зависимости от части речи слово может видоизменяться по нескольким из следующих категорий [Spurkland, 1989]:

- падеж (именительный, винительный, дательный и родительный);
- число (единственное и множественное, для личных местоимений также двойственное: *þú, (þ)it, þér*; отметим, что современный исландский, хотя и не сохранил трихотомии личных местоимений, использует форму двойственного числа вместо множественного: *þu — þið*);
- род (мужской, женский и средний);
- определенность, в отличие от среднеанглийского, реализуется морфологически (точнее, агглютинативно), в том числе для существительных. У прилагательных, как и в среднеанглийском, наблюдаются сильное и слабое склонения с привязкой к определенности подчиняющего существительного;
- степень сравнения (положительная, сравнительная и превосходная);
- лицо (первое, второе и третье);
- время (настоящее и прошедшее, будущее выражается только модально);
- наклонение (изъявительное, сослагательное и повелительное);
- аспект (регулярный и перфективный);
- залог (действительный, страдательный и медиопассив, примерно соответствующий по смыслу возвратному глаголу: *hann lagði**sk** í rekkju*, «он лег (положил **себя**) в кровать», ср. нем. *Er lagte **sich***. Примечательно, что медиопассив сохранился в современных скандинавских языках как

синтетический пассив, альтернативный обычной форме с причастием: норв. *dataen blir fjernet — dataen fjernes*.

Отметим, что предлоги, союзы, междометия и количественные числительные не имели никаких формоизменений.

Отдельной группой выделяют детерминативы: притяжательные и указательные изменения, а также кванторы (такие слова как *allr, sumr, hverr*, т.е. «все», «некоторые», «каждый»). Их общим свойством является использование формы, аналогичной категориальным реализациям прилагательных, в т.ч. склонения по падежу, числу и роду; но в отличие от прилагательных, детерминативы не имеют степени сравнения.

В целом картина склонения/спряжения выглядит следующим образом.

Склонение существительных: сильное и слабое.[2] Сильное встречается реже и отличается от слабого большим разнообразием грамматических суффиксов, в том числе инициирующих мутацию корня, гласная в котором меняется по i- или u-умлауту. Слабое склонение, хотя и может использовать умлаут корневой гласной, образует непрямые падежные формы с практически идентичным окончанием. Хеуген приводит следующий пример [Haugen, 1995, p. 110]:

Таблица 1.

Пример сильного склонения существительного

Форма	fjǫrðr	fjarðar	firði	fjǫrð
Тип формы	Ед.ч, им.п.	Ед.ч., р.п.	Ед.ч., д.п.	Ед.ч., в.п.
Процесс	u-умлаут		i-умлаут, выпадение /j/	u-умлаут

Степенное склонение. Образование сравнительной и превосходной степени у прилагательных проходило по трем типам: суффиксальному (прибавление окончаний -ar, -ast, напр., *spak – spakar – spakast*, «слабый»), суффиксально-фонетическому (прибавление окончаний -r, -st + i-умлаут в корне,

[2] Здесь и далее термины «сильные» и «слабое склонение» употребляются безотносительно склонений, выделяемых по фонологическим типам основ.

напр., *lang – lengr – lengst*, «длинный»), а также в редких случаях суффиксально-супплетивному (прибавление окончаний -r, -st + изменение основы слова, напр., *ill – verr – verst*, «плохой»). Отметим, что не у всех прилагательных имелась полная парадигма; так, прилагательные, обозначавшие относительное положение в пространстве и времени, как правило, имели только формы сравнительной и превосходной степени: *efr/efst*, «позднее/самый поздний».

Склонение прилагательных: сильное и слабое. По морфофонетическим принципам, отнесенным к склонению прилагательных, также видоизменялись детерминативы и причастия. Сильному склонению могли следовать прилагательные в положительной и превосходной степенях сравнения, причастие совершенного вида («партицип перфектум») как сильных, так и слабых глаголов, а также притяжательные, указательные и количественные (кванторы) детерминативы. Слабое склонение охватывало прилагательные в любой степени сравнения, причастие несовершенного вида («партицип презенс») сильных и слабых глаголов, а также некоторые детерминативы и любые порядковые числительные за исключением *annarr* («второй»), которое всегда склонялось по сильной парадигме. Отличие между парадигмами было корневым: по наличию или отсутствию фонетических процессов в корне.

Местоименное склонение. Местоименное склонение охватывало все типы местоимений, в т.ч. все, отнесенные к детерминативам, при этом полнота парадигмы разнилась от вида к виду. Наиболее полной парадигмой отличались личные местоимения: они склонялись по падежам и **трем** числам (кроме третьего лица); третье лицо также отличалось по родам. Возвратное местоимение «себя» (*sín*) относят к третьему лицу, отмечая, что оно не имело формы именительного падежа. Интересно, что формообразование в третьем лице было синтетическим, во втором и первом – супплетивным. Кванторы и указательные местоимения склонялись по падежам и родам; местоимение «который» (*hvat* и его производные) – по падежам, имело только единственное число.

Спряжение: сильное, слабое и смешанное. Следуя общепринятой для германских языков системе, древнескандинавский оперировал обширным

инструментарием глагольных форм, которые принято делить на финитные (настоящее и прошедшее время во всех наклонениях) и нефинитные (начальная форма и причастия). По сути, только финитные формы следовали глагольному спряжению, меняя свои формы по сильной парадигме, т.е. прибегая к аблауту корневой гласной, ср. *brenna* («жечь» или «гореть») в инфинитиве и *brann* в ед. числе прош. времени изъявительного наклонения; слабой парадигме, т.е. образуя прошедшее время только зубными суффиксами -d, -t, -ð, напр. *kenna – kenndi*, «знать»; и смешанной парадигме, которая задействует оба процесса, напр., в претеритно-презентных глаголах – 10 глаголах, по большей части модальных, которые в настоящем времени склонялись по парадигме претеритума сильных глаголов, но в прошедшем формоизменялись аналогично слабым, например, *skulu – skal – skyldi*, «быть должным». В том, что касается причастных форм глаголов, см. «склонение прилагательных» [Faarlund, 2008].

Отметим, что описанные выше характеристики являются обобщенными; уже в начале нового тысячелетия грамматическая реализация начала расходиться таксономически; так, в норвежской и шведской диалектных группах после 1100 г., т.е. ближе к окончанию древнескандинавского периода, наблюдается редукция падежей с полной ассимиляцией винительного и именительного, с последующим переходом дательного из синтетической в полностью аналитическую форму [Trosterud, 2001].

Среднеанглийский язык — совокупность диахронических вариантов английского языка, охватывающих временной период с Битвы при Гастингсе (1066 г.) до коронации Генриха VIII (1509 г.) [Millward, Hayes, 2012, p. 148]. Такое деление само по себе обусловлено спецификой влияния истории Англии на ее язык: приход Вильгельма Завоевателя к власти по итогам сражения при Гастингсе ознаменовал практически одномоментный переход высших деловых, властных и церковных кругов страны от практически «чистого» германского древнеанглийского языка, содержавшего минимум заимствований (в основном обусловленных взаимодействием с церковью и ее каноничными текстами на греческом и латыни) к англофранцузскому языку – смеси английского языка и

норманнского диалекта французского, на котором говорили завоеватели; с исторической же точки зрения именно тогда Англия перешла к феодальному строю, что укрепило влияние языка знати [Иванова, Чахоян, Беляева, 2006, с. 20], фактически вытеснив английский язык из официального и церковного обращения. Прошло полтора столетия, прежде чем в стране был впервые вновь опубликован национальный законодательный документ на ее родном языке: Magna Carta Libertatem, Великая хартия вольностей 1215 г. [Stubbs, Davis, 1921]. Разумеется, английский язык находил отражение в письменной форме и до Хартии, в частности, в манускриптах переводов библейских и религиозных текстов, таких, как «Проповеди Веспасиана», на народную речь, в различных локальных хартиях [Horobin, Smith, 2002]. В то время как древнеанглийский еще использовал руническую запись (алфавит, именуемый «футарком» по названиям первых рун), в среднеанглийском периоде латиница осталась фактически единственным вариантом письменности, при этом сохранились некоторые буквы рунического происхождения [Ralph, Elliott, 1963]:

- торн (þ) – соответствует глухому межзубному фрикативу;
- эт (ð) – соответствует звонкому межзубному фрикативу;
- йох (ȝ) – соответствует неогубленной аппроксиманте либо звонкому велярному фрикативу, либо глухому палатальному/велярному фрикативу;
- вюнн (ƿ) – соответствует огубленной аппроксиманте.

С точки зрения корпусной лингвистики, среднеанглийский в первую очередь интересен и сложен своим письмом. Отсутствие некоего «стандарта», кодифицированного письма, обусловило существование еще в древнеанглийском как минимум четырех различных систем записи, соответствующих западно-саксонскому, древнекентийскому, древнемерсийскому и древненортумбрийскому диалектам [Hogg, 2002]. Некоторая стандартизация письма (отраженная, главным образом, в текстах Чосера) отмечается к конце среднеанглийского периода и основывается первоочередно на лондонском диалекте, что обусловлено, с одной стороны, восстановлением статуса

английского как национального языка, в том числе языка деловой документации, с другой – централизацией страны и усилением роли столицы.

С точки зрения частеречной структуры среднеанглийский идентичен древнескандинавскому, но имеет некоторые парадигматические и морфологические отличия. В целом для его орфографии и фонетики[3] характерны следующие особенности [Brunner, Johnson, 1970; Horobin, Smith, 2002]:

- взаимозаменяемость графем *u*, *v* при обозначении соответствующих звуков;
- взаимозаменяемость *u* и *w* в диграфах с гласными звуками: *au* = *aw*, *ou* = *ow*, *eu* = *ew* (а также *iu*, *iw*);
- взаимозаменяемость буквы «йог» с *i*, обусловленная выпадением конечных звуков /g, x/ с последующей дифтонгизацией либо удлинением предшествующих гласных, ср. древнеангл. *dæg* и среднеангл. *dæi*, «день». У отдельных писцов — также взаимозаменяемость с g;
- взаимозаменяемость графем y, i;
- непоследовательное удвоение графем e, o, a для обозначения долготы соответствующих гласных;
- вокалические процессы: i-умлаут (протогерм. *manniz* и среднеангл. *men*), в том числе с морфологической (форморазличительной функцией), а также преломление гласных перед кластером *lf/lv* (ср. древнеангл. *seolfor* и среднеангл. *silver*) и аблаут как формообразовательный инструмент в сильном глаголе;
- неоднородное обозначение долготы огубленного гласного заднего ряда: o, ou, u;
- неоднородность диграфов *æ*, *ea* в диахронии: изначально обозначали соответствующие им дифтонг(оид)ы, оба были заменены на *a*, *e* во второй половине среднеанглийского периода в связи с фонетическим развитием, при этом *ea* нашел применение в качестве графемы для звука /ø/.

[3] Рассматриваются единым списком в связи с тем, что в значительной мере орфографические особенности были обусловлены именно непоследовательными хронологически и географически фонетическими изменениями.

- идентичность дифтонга /ai/ на письме: *ai, ay, ei, ey*;
- дистрибуционно-зависимое обозначение звука /x/: *gh* перед *t* либо финально, *ȝ* в остальных случаях;
- синкопирование (выпадение) безударных гласных в частицах и артиклях: *theffect*, *tamende* (= *to amende*);
- синкопирование безударных гласных медиально при изменении словоформы: *days* и *dayes*, *yfel — yfle*;
- разрозненность консонантных кластеров на письме ввиду неравномерного и долговременного перехода согласных звуков, например, *sc* и *sk.*
- влияние старофранцузской орфографии: *nacioun*, «нация».

Французское влияние не обошло стороной и грамматику языка, точнее, морфологию, что выразилось заимствованием многочисленных деривационных морфем, таких, как *-ness, -ity* [Dalton-Puffer, 1996].

С точки зрения грамматики в среднеанглийском языке, особенно в конце XII века и позднее, наблюдается значительное упрощение по сравнению с древнеанглийским языком: четырехпадежная система существительных постепенно деградирует до трехпадежной за счет слияния именительного и винительного падежей; к концу XIII века из всех диалектов, кроме юго-восточных, утрачивается родовое различение; сокращается число различных склонений существительных; полностью исчезает различие сильного и слабого склонения прилагательных. Несмотря на то, что у существительных сильное склонение отчасти сохранено, многие слова ввиду гиперкоррекции переходят в слабое склонение, ср. древнеангл. *stān – stānes* и среднеангл. *stone stonen* (различие типов склонения на сильные и слабые в отношении исторических форм английского языка производится по окончанию *-en* у слов слабого склонения в большей части непрямых падежных форм, при этом умлаут как формообразовательный инструмент задействован слабо) [Расторгуева, 2003].

Сравнительная и превосходная степени прилагательных парадигматически унаследованы из древнеанглийского и идентичны степеням сравнения в

древнескандинавском, т.е. используют также три типа формообразования: суффиксальный, суффиксально-фонетический и суффиксально-супплетивный, при этом примыкание суффикса сравнительной степени к финальному согласному звуку приводит к естественной редукции долгой корневой гласной, а после утраты среднеанглийском финали -e в прилагательных слогообразующий звук /r/ переходит в «беглый» шва-звук, например: *greet – gretter*, «большой»; этот звук сохраняется в превосходной степени. По i-умлауту образуются степенные формы таких прилагательных, как *strong* (*strenger, strengest*), «сильный»; супплетивно – *ille* (*werse, werst*), «плохой». Развивается зачаточное в древнеанглийском языке образование степеней сравнения перифразом через *mo, more, most* [Mayhew, Skeat, Everson, 2011]. Отметим, что у среднеанглийского активно развивается такое явление, как субстантивация прилагательных: *þe innocent*, «невинные (люди)» [Mustanoja, Gelderen, 2016]. Родовое и падежное склонения редуцируются, в частности, прилагательное лишается существовавшей в древнеанглийском языке формы творительного падежа (сливается с дательным) и винительного (сливается с именительным). Предикативные прилагательные, в отличие от древнескандинавского, вовсе не имеют флексии; также не меняют своей формы прилагательные, заимствованные из французского, за исключением тех случаев, когда вместе со словом была заимствована какая-либо его альтернативная форма, например, множественное число: *weyes esperituels*, «духовные пути». По сути, адъективальная парадигма сводится к различию формы числа, общего и среднего родов.

Прилагательные становятся частым источником образования наречий за счет суффиксации через *-liche, ly* – адвербиальных суффиксов, развившихся из древнеанглийских *-liċe, ligr*. Так же, как прилагательные, склоняются порядковые числительные, причем следуя слабому склонению за исключением *oþer*, «другой», «второй», что идентично древнескандинавскому языку. Из количественного числительного *ān*, «один» со временем развивается неопределенный артикль, в то время как определенный формируется из указательного местоимения *þat*. В отличие от древнескандинавского,

определенность существительного не выражается суффиксально и не влияет на склонение прилагательного ввиду слияния сильного и слабого склонения у последних [Horobin, Smith, 2002].

Местоимения, как и в древнескандинавском языке, делятся на личные, притяжательные, указательные и количественные (кванторы), причем последние три также относят к детерминативам, включающим в том числе артикли [Mustanoja, Gelderen, 2016]. При этом у личных местоимений наблюдается значительная редукция парадигмы до двух падежей (именительного и косвенного), хотя некоторые источники все еще различают четыре германских падежа, признавая при этом совпадение форм в большей их части, например, *ich — me — my — me* (обратим внимание, что генитив *my* фактически стал альтернативой притяжательному местоимению). Отсутствует форма двойственного числа. Стоит отметить, что для среднеанглийского языка характерно клитическое употребление личных местоимений и глагольной частицы: *wilte = wilt þu; nele = ne wille* [Haeberli, Ingham, 2007]. Род местоимения перестает быть грамматическим и фактически определяется только по семантике: при отсылке к неодушевленному субъекту используется средний род. Отдельно рассматриваются относительные (*which*) и вопросительные местоимения.

Если говорить о глагольной системе, то значительная часть глагольных классов утрачивается, сливаясь с другими, при этом, по сути, остаются лишь три категории глаголов: сильные, слабые и неправильные; спряжение производится по лицу, числу, времени и наклонению. Сильный глагол образует форму прошедшего времени по аблауту: *binden – bounde*, «связывать»; слабые – дентальным суффиксом *-d*: *louen – louede*, «любить»; неправильные глаголы полагаются на супплетивную формацию: *ben – was/were*. Имеется также набор из 11 претеритно-презентных глаголов, сохранившихся со времен древнеанглийского языка, например, *schal – scholde*. Сохраняется деление на настоящее и прошедшее время, однако добавляется перфективный аспект, образуемый, как и в древнескандинавском языке, с помощью вспомогательного глагола и причастия; также используется система трех наклонений:

изъявительного, сослагательного и повелительного. Будущее время не выражено морфологически и вместо этого реализуется модально с помощью глагола *schal*; несмотря на то, что причастие настоящего времени уже присутствует как глагольная форма, оно следует адъективальной парадигме и употребляется атрибутивно: прогрессивный аспект глагола разовьется только в новоанглийском периоде, хотя на севере Англии и в Шотландии уже отмечаются попытки использовать подобные конструкты с глаголами движения: *þe childe was cumand*, «ребенок шел». Отметим, что на юге подобное причастие использовало окончание *-ing/yng/ung*, смешиваясь морфологически с отглагольным существительным/герундием [Ильиш, 1973].

Рассмотрим все вышеозначенные процессы и особенности с точки зрения того, как они влияют на реализацию поставленной задачи, а именно на машинное обучение частеречной разметке.

Фонетические процессы. В древнескандинавском языке имеются три вида вокальных мутаций: a-, i-, u-умлауты, причем второй и третий несут морфологическую/формообразовательную функцию; в среднеанглийском присутствует только i-умлаут, его морфологическая функция выражена слабее. С точки зрения машинного обучения, речь заходит о необходимости обеспечивать идентичное определение частеречной принадлежности, например, слов *mann/menn*; *tala/tǫlum* при различении их формы (впрочем, так как в задачи текущего исследования не входит обучение морфограмматической разметке, различение алгоритмом словоформ одного слова не является критической функцией); с другой стороны, сами по себе мутации являются частеречно-специфичными признаками, при этом не образуя морфов как таковых, что усложняет реализацию морфологических алгоритмов, так как одним из атрибутов слова становится его корневая гласная, по сути, образующая своего рода инфикс.

Структура частей речи. С одной стороны, идентична в обоих языках, что позволяет использовать один и тот же набор частеречных помет (PoS-tags); с другой – части речи отличаются морфологическими признаками, при этом в

среднеанглийском гораздо раньше наблюдается переход к аналитизму, что несколько нивелирует морфологическое различие между именными частями речи и вкупе с перенесенной из древнеанглийского клитикализацией глагольной частицы и личного местоимения усложняет задачу определения частеречной принадлежности по морфологическому признаку. Еще одну сложность, обусловленную аналитизмом, представляет собой субстантивация прилагательных в среднеанглийском языке, использование глагольного причастия в определительной функции в древнескандинавском. В целях разрешения подобной проблемы считаем оправданной (и вынужденной) мерой идентификацию причастия настоящего времени (несовершенного причастия) в качестве прилагательного ввиду морфологической и функциональной идентичности даже в тех случаях, когда с его помощью производится не атрибуция, а образование глагольной формы в прогрессивном аспекте.

Морфология именных частей речи значительно упрощена в среднеанглийском по сравнению с древнескандинавским, в частности, сокращен набор фонетических основ существительных, глагольных классов, практически отмерло различение сильного и слабого склонения прилагательных. Наибольшую проблему может представлять имя прилагательное с его окончанием *-st* в превосходной степени, которое совпадает с глагольным окончанием второго лица единственного числа в индикативе. В обоих языках значительно разнится формообразование сильного и слабого склонения существительных. Особняком стоит супплетивная формация: в то время как для прилагательных она не представляет особой проблемы, могут возникать сложности с машинной идентификацией личных местоимений, что, на наш взгляд, должно компенсироваться уникальной графической формой таких местоимений и их ничтожно малым количеством.

Глагол. Наиболее богатая и уникальная с точки зрения морфологии часть речи, что в теории должно обеспечить простоту ее дифференциации по графическому представлению; так, в обоих языках применяется уникальная система личных и временных окончаний, в том числе агглютинативно в

прошедшем времени; в обоих языках именно глаголу наиболее свойственно проявление аблаута как формоизменяющего механизма; кроме того, в древнескандинавском, в отличие от среднеанглийского, присутствует строго формальное выражение медиопассива, окончание которого происходит от возвратного местоимения. С другой стороны, наблюдается множественная коллизия глагольных и именных морфов, в частности, совпадение личного окончания второго лица единственного числа и маркера суперлативной формы прилагательных (-st) в среднеанглийском языке ввиду отсутствия ротацизма превосходной степени; совпадение окончаний презенса и компаративной формы прилагательных в древнескандинавском ввиду ротацизма первых; совпадение маркера инфинитива и неименительных форм слабого склонения существительных в среднеанглийском языке *(-en*); схожесть маркера медиопассива во втором и третьем лицах с типовым окончанием начальной формы прилагательного в древнескандинавском языке *(-sk*; стоит отметить сравнительную редкость текстового употребления именно начальной формы, так что данная особенность может и не представлять собой значительной проблемы). Особняком встают нефинитные формы, такие, как причастие настоящего и прошедшего времени (иначе — несовершенное и совершенное причастия, первое и второе причастия), использующие парадигму прилагательных (с этой точки зрения в среднеанглийском языке вообще отсутствует синтетически выраженный пассивный залог).

Графика. Оба языка отчасти задействуют символы рунического происхождения. Ввиду неоднородности письма, следующего за диалектными различиями (хотя и демонстрирующего некоторое стремление к «столичной» записи), а также под влиянием французского языка в среднеанглийском наблюдается значительный разнобой в графической фиксации текста, что усложняет задачу его кодирования: с одной стороны, необходимо обеспечить идентичное или практически идентичное представление, например, слов *scilde* и *skylde*, «щит»; с другой – используемая репрезентация не должна приводить к смешению таких слов с практически омографичными лексемами иных частей,

например, *scholde*. Решение данного вопроса является одним из наиболее сложных аспектов поставленной задачи.

Таким образом, разрабатываемая методика машинного обучения должна обеспечивать адекватное различение морфологических схожих именных частей речи, реализуемых практически идентичными морфами отдельных форм прилагательных и глаголов, устранение коллизии частеречного определения per se, будучи при этом толерантной к некоторому орфографическому расхождению, особенно применительно к среднеанглийскому тексту.

1.3.1. Методы векторизации текста, применимые к историческому тексту

В задачах машинного обучения одной из первостепенных решаемых проблем является представление данных на входе алгоритма, так как в условиях неизменности самих алгоритмов (в конечном счете, та же модель случайного леса сама по себе не отличается в задачах по прогнозированию причин сбоя работы SSD-хранилищ в центрах обработки данных [Narayanan и др., 2016] и в лингвистических задачах [Gries, 2015]), именно репрезентация данных определяет то, насколько эффективно алгоритм справится с возложенной на него задачей [The importance of Data Representation - Data Mining and Reporting]. В этой связи представляется важным проанализировать распространенные в области обработки текста на естественном языке способы машиночитаемого представления данных и определить, насколько эффективным будет приложение того или иного метода в задачах настоящего исследования.

Так как алгоритмы машинного обучения на вход получают векторы (в данном случае под векторами понимаются последовательности из n-вещественных чисел, по сути – точки в n-мерном пространстве), машиночитаемое представление текста переводит его в векторную форму. Единицей текстовых данных обычно принимается слово, таким образом, каждому слову соответствует один вектор, причем векторы схожих в том или ином аспекте слов должны располагаться близко друг к другу в их пространстве

над полем вещественных чисел. Методы такого представления, как правило, в той или иной степени полагаются на распределительную гипотезу Фирта: «Понимание значения слова приходит из его окружения» (англ. You shall know a word by the company it keeps) [Firth, 1957]; в целом распространенные методы векторизации делятся на статистические, такие, как анализ латентной семантики, и предиктивные, например, нейронные вероятностные языковые модели. Популярны такие модели, как «непрерывный мешок слов» (англ. continuous bag of words) и skip-gram [Mikolov и др., 2013]. В то время подобные модели хорошо выполняют свои функции в задачах информационного поиска, индексации и фильтрации текстовых данных, обеспечивая, например, определение псевдо-синонимов (используется поисковыми системами для выдачи результатов, схожих семантически, но не идентичных запросу пользователя), одним из ранее сформулированных требований к векторизации является ее способность кодировать значимую для частеречной классификации внутреннюю информацию, содержащуюся в слове, а именно – морфемы и их графическую реализацию, так как именно морфемы являются наиболее значимыми маркерами частеречной принадлежности в германских языках; векторное представление слова в контексте предназначено для анализа семантики, но не морфографемики. В этой связи необходимо обратиться к средствам компьютерной морфологии и используемым в них алгоритмам морфологического кодирования.

Одним из способов такого кодирования является описанный в работе [Takala, 2016] метод векторизации «основа+окончание» – сублексическая модель, в рамках которой слово разбивается на две части, каждая при этом кодируется собственным вектором; таким образом, получаем два векторных пространства: пространство основ и пространство суффиксов. Проблема данного подхода применительно к среднеанглийскому и древнескандинавскому языкам очевидна: в обоих языках присутствует как конечная морфология, так и внутренняя (более того, у некоторых слов присутствует только последняя), например, аблауты и умлауты в морфологической функции; кроме того, среднеанглийский язык формирует глагольное причастие префиксацией с *ge-*

или иными приставками. Таким образом, в дополнение к основному векторному пространству, кодирующему основы, необходимы как минимум три добавочных пространства, кодирующие префиксы, суффиксы и инфиксы (ранее в работе отмечалось, что мутация корневой гласной как средство формообразования может считаться инфиксом). Кроме того, само выделение морфов в целях формирования таких пространств становится отдельным шагом всего алгоритма обучения, автоматизация которого осложнена. Одним из способов избежать необходимости ручного формирования перечней морфов является использование моделей морфологической сегментации, например, Morfessor, которые успешно применяются в отношении современных морфографемически сложных языков, таких, как русский [Sadov, Kutuzov, 2018]. Попытка использовать Morfessor 2.0 [Smit и др., 2014] в рамках настоящей работы для морфологической сегментации текста на английском языке показала его сравнительно низкую эффективность, возможно, обусловленную малым размером выборки. Так, в тексте «Проповедей Веспасиана» при обучении на неразмеченном тексте всей диахронической части Хельсинкского корпуса неидентифицированы глагольные префиксы, поделены некоторые местоимения и предлоги, ничего, кроме корневых морфем, не содержащие; при этом корректно выделены маркеры падежных форм большинства существительных и прилагательных, что, однако, недостаточно для реализации подобного кодирования.

Поскольку морфологический маркер, как правило, использует идентичный набор символов даже в условиях неустойчивости орфографии, при этом имеет фиксированное положение в слове, представляется осмысленным применение моделей посимвольной векторизации. Обоснованность подобного решения подтверждается в работе [Cao, Rei, 2016], где заявляется, что модели на уровне символов позволяют повысить точность и полноту частеречной классификации. Там же упоминается модель Char2Vec, задействующая алгоритм skip-gram и простую нейронную сеть в целях отображения каждого символа из алфавита, встречающегося в корпусе текста, на двухмерном пространстве. Так

обеспечивается схожесть кодирования символов, возникающих в схожих контекстах (т.е. имеющих схожее лингвистическое окружение) [Шенбин, 2017], что наглядно показано на рис. 2, где, например, видна группировка гласных, что говорит о схожести их дистрибуции.

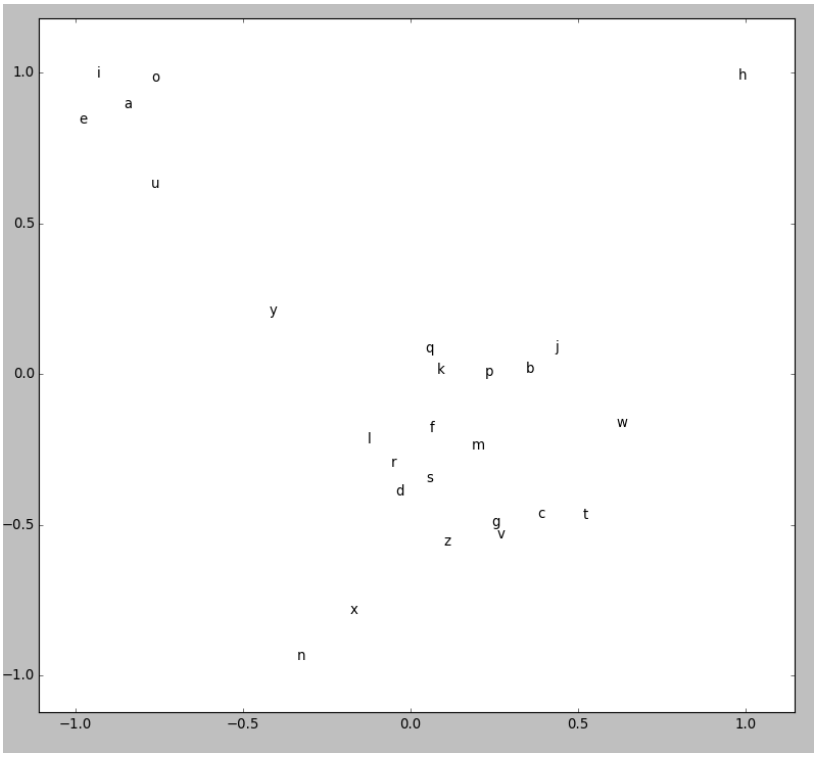


Рис. 2. Двухмерное отображение вывода модели Char2Vec

Хотя модели посимвольного кодирования, аналогичные Char2Vec, успешно применяются в задачах обучения частеречной разметке вкупе с моделями пословной векторизации, например, в отношении корпуса португальского языка [Nogueira, Santos, Zadrozny, 2014], подобное кодирование требует предварительного обучения самой векторизующей модели на достаточно большом объеме текстовых данных (в указанном исследовании речь идет о корпусе португалоязычной Википедии); кроме того, подобное обучение является ресурсоемким и, как правило, задействует серверные мощности [Там же], которые недоступны в рамках настоящего исследования.

Наиболее простой и компактный с точки зрения размерности, а также наименее ресурсоемкий способ посимвольной векторизации слова предлагает П. Такала [Takala, 2016] в виде так называемого «метода скользящего среднего» (англ. moving average approach); суть метода заключается в том, чтобы

сгенерировать вектор фиксированной размерности, равной размеру алфавита. Для этого каждому символу в алфавите назначается определенное измерение, после чего в векторе, представляющем то или иное слово, к значению измерения прибавляется позиционно-зависящий вес встречающихся в нем символов, рассчитанный как скользящее среднее. Так получаем репрезентацию слова w = ($w_a w_b \ldots w_z$),

$$w_a = \sum \frac{(1-\alpha)^{C_a}}{Z},$$

где c — индекс обрабатываемого символа (0 для первого символа в слове, 1 для второго и т.д.), α соответствует гипер-параметру, обусловливающему понижение значения на выходе, Z — нормализатор, значение которого пропорционально длине слова. Данная формула обеспечивает низкое выходное значение в отношении символов, находящихся в конце слова, по сравнению с теми, которые употреблены в начале; таким образом, метод скользящего среднего эффективно кодирует не только сам факт употребления символа в слове, но и его положение относительно других символов, что может быть достаточно для целей настоящего исследования. Процедура повторяется в обратном порядке (например, слово — оволс), затем производится простой подсчет символов; полученные таким образом три вектора конкатенируются.

П. Такала в своей работе заявляет о высокой эффективности последнего метода по сравнению с простой векторизацией слов применительно к материалу корпуса финноязычной Википедии [Там же], хотя не уточняет характер решаемой задачи. С учетом особенностей языкового материала, изложенных в предыдущем пункте, считаем такую репрезентацию достаточной, так как она обеспечивает эффективное различение позиций в векторном пространстве слов, использующих разные символы на одинаковых позициях внутри лексемы, при этом две реализации одного и того же слова, различающиеся только одним символом (что характерно для среднеанглийского языка), оказываются на схожих позициях, что должно обеспечить их одинаковую классификацию.

1.3.2. Методы машинного обучения, применимые к историческому тексту

Выбранный ранее метод векторного представления слов позволяет получить на входе алгоритма машинного обучения набор векторов, сопровождаемых частеречными пометами, при этом сами векторы несут информацию только о составе и расположении символов в слове, но не о лингвистическом окружении (контексте). Из различных моделей машинного обучения, способных обрабатывать подобный материал, выделяется несколько: метод k ближайших соседей, относящийся к т.н. «ленивому обучению» (англ. lazy learning), метод опорных векторов, наивный алгоритм, основанный на теореме Байеса (англ. Naïve Bayes), и логистическая регрессия, относящиеся к линейным классификаторам, модель случайного леса (англ. random forest model) и деревья принятия решений, в т.ч. усиленные (англ. boosted), а также многослойный перцептрон, являющийся простой нейронной сетью обратного распространения. Отметим, что все указанные алгоритмы являются классификаторами, что соответствует характеру поставленной задачи, являющейся, по сути, проблемой классификации – одной из классических задач в области обработки естественных языков [Найденова, Невзорова, 2008]. Рассмотрим эти модели по отдельности.

Самым простым из всех указанных алгоритмов является kNN, или метод k ближайших соседей. Данный алгоритм основан на прямом измерении расстояний между векторами по какой-либо метрике, причем наиболее часто используется т.н. «евклидово» расстояние, рассчитываемое по формуле:

$$|\vec{v} - \vec{w}| = \sqrt{\sum_{i=1}^{2}(v_i - w_i)^2},$$

т.е. евклидово расстояние между векторами $\vec{v}$ и $\vec{w}$ рассчитывается как модуль корня из суммы квадратов разностей их соответствующих координат. Классификатор работает за счет нахождения *k* ближайших к классифицируемому вектору соседей в данных модели и определения их классов [Aha, Kibler, Albert, 1991], после чего искомый класс назначается простым голосованием большинства, что хорошо показано на следующей иллюстрации:

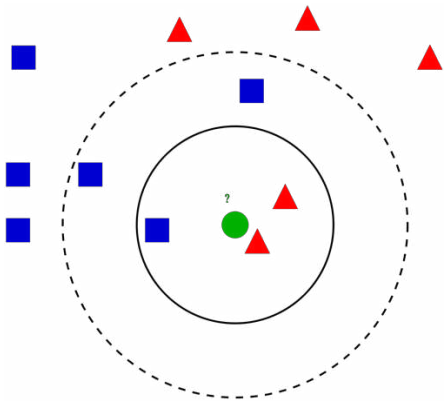


Рис. 3. Принцип работы «ленивого» классификатора kNN

На рисунке видно, что метод определит класс неизвестного объекта в центре как «красный треугольник», если k = 3, как синий квадрат, если k = 5, как красный треугольник, если k = 11. Число k обычно принимается нечетным, чтобы обеспечить большинство в случае бинарной классификации. Несмотря на свою простоту, данный алгоритм эффективно применяется в задачах классификации текстовых документов [Guo и др., 2004], также показал свою эффективность в рамках предсказания позиции ударного слова в словах нидерландского языка [Daelemans, Durieux, Gillis, 1994].

Несколько сложнее принцип работы опорных векторов – алгоритма, предназначенного в первую очередь для бинарной классификации, но также способного производить категоризацию по большему числу классов за счет деления общей задачи на кластер бинарных подзадач, где один класс противопоставляется всем остальным [Cristianini, Shawe-Taylor, 2000]. Алгоритм основывается на поиске т.н. «оптимальной гиперплоскости» – плоскости (линии, в случае двухмерного пространства), делящей векторное пространство так, чтобы с каждой «стороны» от нее находились векторы только одного класса из обучающей выборки, при этом расстояние до ближайших из них (т.н. «опорных» векторов) было максимально возможным. Принцип работы хорошо показан на следующем рисунке:

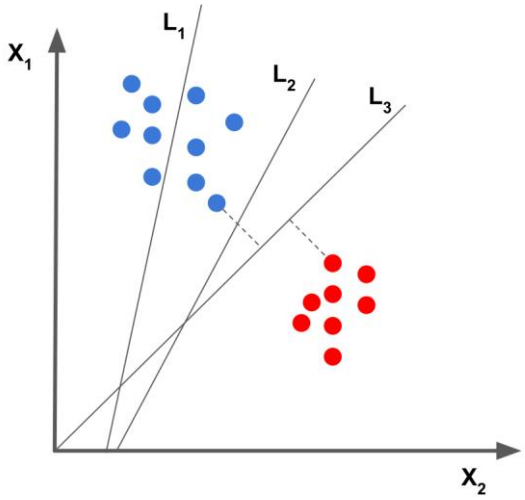


Рис. 4. Принцип работы опорных векторов

Плоскость L_1 не является оптимальной, так как не обеспечивает деление векторного пространства с вытеснением векторов одного класса строго «по одну сторону» от себя; плоскость L_2 также не является оптимальной, поскольку не максимизирует расстояние от себя до опорных векторов; плоскость L_3 является оптимальной гиперплоскостью, поскольку адекватно разделяет векторы, содержащиеся в данном пространстве, на классы, при этом расстояние до опорных векторов становится максимально возможным.

Метод опорных векторов успешно применяется в решении различных лингвистических задач, таких, как классификация текстов [Leopold, Kindermann, 2002; Manning и др., 2012], а также определение их авторства [Diederich и др., 2003], однако нам не удалось найти сведений об успешной реализации подобной модели в задачах автоматизации частеречной разметки.

Теорема Байеса в ее базовой форме достаточно проста:

$$P(A|B) = \frac{P(A)P(A|B)}{P(B)},$$

т.е. вероятность события *A* при условии наступления события *B* определяется как частное произведения вероятности наступления события *A* per se, вероятности наступления события *B* при условии наступления события *A* и вероятности наступления события *B* per se [Hartshorn, 2016].

Алгоритм, использующий теорему Байеса в ее базовой форме, называют «наивным»; в его основе лежит строе предположение независимости. В дополненном и расширенном виде байесовы модели могут применяться в том

числе для решения лингвистических задач, таких, как моделирование секвенциального восприятия текста человеком [Narayanan, Narayanan, Jurafsky, 1998] или определение авторства [Mosteller, Wallace, 1963]; кроме того, подобные модели использовались в том числе для решения задач частеречной разметки [Hasan, Ng, 2009].

Достаточно популярным средством классификации является логистическая регрессия – сигмоидная функция, предсказывающая вероятность того, что y = 1 при заданном значении x. Логистическая регрессия основана на логистическом уравнении:

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}},$$

где p – вероятность, β_0 – значение зависимой переменной y при $x = 0$, $\beta_1 x$ – изменение зависимой переменной при x=1. В качестве степенного показателя используется универиантная модель регрессии. В отличие от линейной регрессии, логистическая обеспечивает эффективный захват «выбросов» – значений, находящихся аномально далеко от прочих показателей в выборке. Иллюстративно такая регрессия обычно демонстрируется графиком сигмоиды:

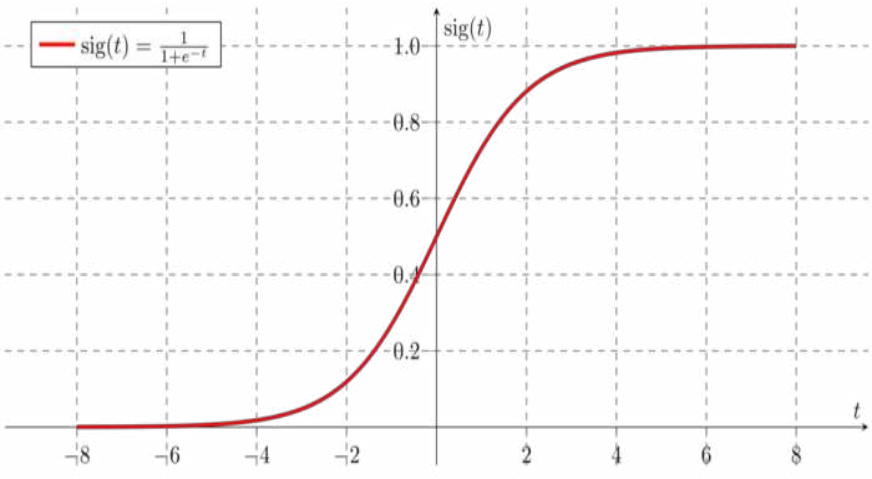


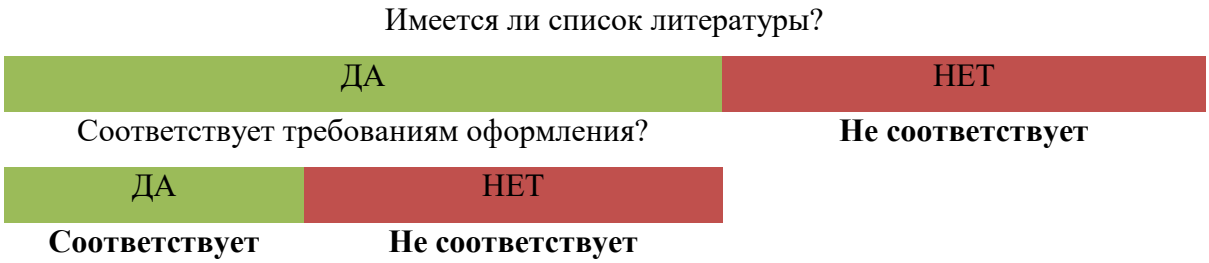
Рис. 5. Сигмоида

Логистическая регрессия является одним из важнейших инструментов машинного обучения и находит применение в большинстве задач обработки текста на естественном языке [Jurafsky, Martin, 2009]. Одной из таких задач является категоризация текста [Genkin, Lewis, Madigan, 2007], его анализ [Taddy,

2013]; применима логистическая регрессия также в отношении частеречной разметки [Pranckevicius, Marcinkevicius, 2016], [Cellier и др., 2016].

Популярным средством классификации являются деревья принятия решений (англ. decision trees). Алгоритм часто визуализируется в виде простой последовательности логических шагов в форме ответов на вопросы, необходимые для принятия того или иного решения, например:

– Соответствует ли статья требованиям к оформлению списка литературы?



В соответствии с такой визуализацией дерево принятия решений часто использует булевы функции, аргумент (x) которых принимает только два значения: 1 (да) или 0 (нет). При обучении в «корне» дерева помещается некоторый выражаемый численно атрибут Q, или признак объектов из обучающей выборки (назовем ее обучающим множеством A); при росте и ветвлении дерева осуществляется перенос на ветви только тех объектов из множества A, в отношении которых значение выбранного атрибута равно какому-либо из возможных значений (в случае булевых функций диапазон значений сводится к бинарной дихотомии). Основной алгоритм выбора атрибута в корне основан на принципе прироста информации и информационной энтропии — метрики информационной полноты выборки. Так, если выборка однородна, то энтропия равна 0, если делится на классы поровну, то 1. Прирост информации (также называемый информационным выигрышем, англ. information gain) рассчитывается следующим образом:

$$\mathrm{Gain(A,Q)} = \mathrm{H(A,S)} - \sum_{i=1}^{q} \frac{|A_i|}{|A|} \mathrm{H(A_i,S)},$$

где A_i — множество элементов A, у которых атрибут Q имеет значение i, а $H(A,S)$ — информационная энтропия множества A в отношении свойства S его атрибутов, определяемая как

$$H(A,S) = -\sum_{i=1}^{s} \frac{m_i}{n} log \frac{m_i}{n},$$

где m_i — число случаев, в которых у элементов данном множестве A реализуется некое свойство S, n — размер данного множества. На каждом шаге алгоритм выбирает такой атрибут Q, при котором информационный прирост будет максимален.

Отметим, что такой принцип выбора атрибутов для ветвления дерева принятия решений используется алгоритмом ID3 и не является единственно возможным, однако пользуется популярностью ввиду своей вычислительной простоты и малой ресурсоемкости [Шестаков, 2012].

Развитием принципа деревьев принятия решений является модель случайного леса – алгоритм, «выращивающий» множество подобных деревьев случайным образом и задействующих разные атрибуты, выбирая их по разным принципам (например, с помощью коэффициента Джини), а затем определяющий окончание решение классификатора методом простого либо взвешенного голосования полученных деревьев [Breiman, 2001]. Случайный лес как средство классификации активно используется в лингвистических задачах, применяющих машинное обучение, например, в целях моделирования фонетической редукции на материале корпуса записей устного английского языка [Dilts, 2013], в рамках преобразования текста в речь в китайском языке [Oparin, Lamel, Gauvain, 2010]; однако Грис и др. призывают проявлять осторожность в применении классификационных деревьев и случайного леса в корпусной лингвистике [Gries, 2015].

Последним из указанных ранее алгоритмов является многослойный перцептрон – простая нейронная сеть обратного распространения [Du, Swamy, 2014]. Перцептрон – сеть, получающая на входе параметры x_1, x_2… x_n и их соответствующие весовые коэффициенты w_1, w_2…w_n; далее производится расчет функции активации, которая перемножает значения входных параметров и их весовых коэффициентов, производя выходное значение y. Работа перцептрона описывается следующим уравнением:

$$y = \phi(\sum_{i=1}^{n} w_i x_i + b),$$

где w — вектор весовых коэффициентов, x — вектор входных значений, b – систематическая ошибка, возникающая ввиду наличия ошибочных допущений в обучающем алгоритме и приводящая к «недостаточной подгонке» (англ. underfitting) значений; определяется как разность между реальным значением оцениваемого параметра и значением, предсказанным по итогам выполнения алгоритмической оценки.

Многослойный перцептрон также применяет как минимум один скрытый слой между входом и выходом модели, задействуя метод обратного распространения в качестве обучающего алгоритма. При прямом проходе сигнал идет от входного слоя через скрытые на выход, после чего выходное значение сравнивается с фактическими значениями, данными в обучающей выборке; так рассчитывается ошибка. При обратном проходе осуществляется дифференциация, производящая градиент ошибки, после чего весовые коэффициенты корректируются так, чтобы обеспечить соответствие вывода у фактическим значениям в выборке; алгоритм воспроизводится до тех пор, пока дальнейшее понижение ошибки не станет невозможным; такое состояние называют схождением.

Проиллюстрировать работу многослойного перцептрона можно с помощью следующей схемы [Sonawane, Patil, 2014]:

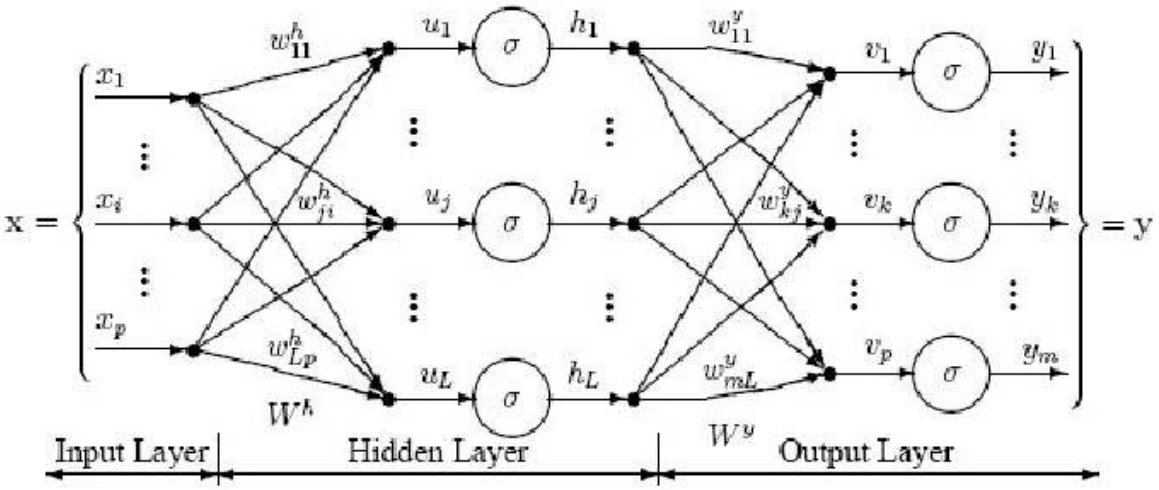


Рис. 6. Работа многослойного перцептрона. h_n соответствуют скрытым слоям, сигма — исполнению сигмоидной функции

Одним из первых приложений многослойного перцептрона в лингвистических задачах стала система NetTalk, обеспечивающая фонетическую транскрипцию машиночитаемого текста на английском языке [Hemming, 2003]. И. Трипати, М. Оукс и С. Вермтер обратились к многослойному перцептрону при составлении параллельного многоалгоритмического (ансамблевого) классификатора в рамках задачи категоризации текста [Tripathi, Oakes, Wermter, 2012]. Кроме того, данный тип алгоритмов успешно применяется в задачах обучения автоматических частеречных аннотаторов [Young и др., 2018]. Мы же попытались применить их к историческому тексту.

1.3.3. Методы ансамблирования

Нами ранее было показано, что отдельные методы машинного обучения могут допускать существенные ошибки классификации, при этом сами разновидности ошибок – некорректное присвоение того или иного класса – разнятся от классификатора к классификатору [Karimov, Akinin, Yakymets, 2018]. Теоретически, комбинирование различных алгоритмов может обеспечить компенсацию ошибок отдельных техник машинного обучения, т.е. т.н. «ансамбль» должен оказаться эффективнее отдельно взятых моделей в его составе. В этой связи считаем уместным рассмотреть существующие методы ансамблирования.

Ансамбль представляет из себя мета-алгоритм, сочетающий несколько методик машинного обучения в рамках одной предиктивной модели, что позволяет снизить дисперсию (при использовании «бэггинга»). или «предвзятость» классификаторов (за счет «бустинга»), а также повысить прогностическую точность группы (путем «стекинга»).[4] Методы построения ансамбля можно условно поделить на две основные категории:

[4] Здесь и далее в отношении ансамблевых методов применяется заимствованная англоязычная лексика ввиду отсутствия в русском языке устоявшихся эквивалентов.

- Секвенциальные, или последовательные ансамбли, где базовые классификаторы реализуются последовательно, например, алгоритм AdaBoost. Основополагающим принципом работы таких ансамблей является реализация зависимости между алгоритмами в их составе. Общая эффективность работы и точность классификации повышаются путем присвоения повышенных весовых коэффициентов тем примерам в обучающей выборке, которые ранее были классифицированы некорректно.

- Параллельные ансамбли, в рамках которых все базовые классификаторы обучаются и генерируются одновременно друг с другом; в качестве примера можно привести уже рассмотренный ранее алгоритм случайного леса, представляющий собой параллельный ансамбль случайных деревьев принятия решений, каждое из которых реализуется как отдельных алгоритм. Основополагающим принципом работы таких конструктов является реализация независимых отношений между базовыми методами машинного обучения в их составе и снижения ошибки путем усреднения.

Помимо данной классификации, ансамбли также подразделяются на однородные (гомогенные), состоящие из базовых алгоритмов одного и того же типа, и разнородные (гетерогенные), где базовые классификаторы относятся к разным типам. Повышение эффективности и точности ансамбля по сравнению с его отдельными компонентами требует разнообразия состава и высокой точности самих базовых алгоритмов.

Одним из самых популярных методов ансамблирования является т.н. «бэггинг» (англ. bagging, bootstrap aggregation). Бэггинг обеспечивает понижение дисперсии оценки за счет усреднения результатов работы разных оценивающих алгоритмов. Похожая техника лежит в основе работы моделей случайного леса, где обучается M различных деревьев принятия решений на разных частях выборки (выбираются случайным образом) и рассчитывает итоговый результат по формуле:

$$f(x) = 1/M \sum_{m=1}^{M} f_m(x)$$

Обобщение результата работы ансамбля производится методом голосования при использовании в задачах классификации, усреднения в целях регрессии.

Еще один популярный метод – «бустинг» (усиление, англ. boosting). Принцип усиления заключается в подгонке некоторой последовательности слабых классификаторов, например, малоразмерных деревьев принятия решений, к взвешенному набору данных, при этом весовой коэффициент, присваиваемый примерам из обучающей выборки, повышается у тех объектов, которые ранее были классифицированы некорректно. Итоговый результат рассчитывается методом взвешенного голосования. Отличается от бэггинга тем, что базовые классификаторы обучаются последовательно, с корректировкой весовых коэффициентов в данных.

Самой широко используемой реализацией бустинга является алгоритм AdaBoost (adaptive boosting). Данный алгоритм обучает первый в последовательности базовых классификаторов с применением равнозначных весовых коэффициентов, затем повторяет процесс обучения, наращивая значение таких коэффициентов в отношении всех объектов выборки, которые по итогам первого раунда были классифицированы некорректно, и понижая вес правильно классифицированных объектов, при этом на втором раунде используется уже другой классификатор. Процесс повторяется еще раз с применением третьего классификатора и присвоением высоких весовых коэффициентов тем объектам обучающей выборки, где ранее использованные классификаторы показали неравную точность. В конечном итоге каждому классификатору присваивается взвешенный коэиффициент ошибок таким образом, чтобы более точные классификаторы имели более высокий «вес» при голосовании по итогам исполнения ансамбля, при этом значение веса разнится не только от классификатора к классификатору, но и между различными областями выборки [Tharwat, 2018].

Еще один популярный алгоритм ансамблирования – стекинг; под этим термином понимают комбинирование нескольких моделей классификации, или

регрессии, посредством мета-классификатора, или мета-регрессора. Базовые алгоритмы в составе стека обучаются на полносоставной выборке, затем ансамбль обучает метамодель, на вход которой подаются уже результаты работы базовых классификаторов. Стекинг обычно реализуется в виде гетерогенного ансамбля [Ngo, 2018].

В методе случайных подпространств (англ. random subspace method, RSM) базовые алгоритмы обучаются на различных подмножествах признакового описания, которые также выделяются случайным образом. Этот метод предпочтителен в задачах с большим числом признаков и относительно небольшим числом объектов, а также при наличии избыточных неинформативных признаков. В этих случаях алгоритмы, построенные по части признакового описания, могут обладать лучшей обобщающей способностью по сравнению с алгоритмами, построенными по всем признакам [Гущин, 2015].

Наконец, можно отметить голосование как метод ансамблирования, подразделяющееся на простое и взвешенное, причем фактически простое и взвешенное голосование можно выразить одним и тем же уравнением:

$$b(x) = \sum_{t=1}^{T} w_t \, b_t(x)$$

с той разницей, что при простом голосовании базовых алгоритмов b_1, $b_2 \ldots b_t$ умножается на весовой коэффициент w=1. Как простое, так и взвешенное голосование, в отличие от прочих представленных здесь алгоритмов ансамблирования, ранее применялось в рамках настоящего исследования [Karimov, Akinin, Yakymets, 2018], не обеспечив при этом значимого повышения точности классификации, в связи с чем далее в работе рассматриваться не будет.

ВЫВОДЫ ПО ПЕРВОЙ ГЛАВЕ

В результате изучения теоретических и прикладных трудов по теме исследования было выявлено, что и среднеанглийский, и древнескандинавский языки используют обширный набор морфографемических маркеров, обеспечивающих различение лексем по частеречной принадлежности и реализующихся в тексте диахронического корпуса. Древнескандинавский язык отличается бóльшей морфологической комплексностью, использует значительно более широкий набор подобных маркеров и прибегает к обширному набору фонетических и морфосинтактических средств, отсутствующих в среднеанглийском языке. Так, только в первом из рассматриваемых германских языков присутствуют, помимо i-умлаута, a- и u-мутации, в том числе несущие формообразовательную функцию, а также агглютинативный инструмент маркировки категории определенности в виде суффигированного определенного артикля. Помимо мутации корневой гласной, природа которой лежит в регрессивной ассимиляции под воздействием вокалической финали, присутствует так называемый аблаут – перебой корневой гласной как средство образования формы прошедшего времени сильного глагола, а также некоторых личных форм претеритно-презентных глаголов в обоих языках. Таким образом, речь фактически заходит о существовании в древнескандинавском и среднеанглийском не только суффикса, но и своего рода инфикса; в среднеанглийском также имеется образование одной из нефинитных глагольных форм (второго причастия) префиксом, что само по себе скорее представляет проблему не кодирования текста в векторной форме как таковой, а частеречной классификации самой по себе: второе причастие, будучи нефинитной формой, использует адъективальную парадигму, а не глагольную. Гораздо более значимой проблемой представляется частичное совпадение именных парадигм существительного и прилагательного в их морфологическом выражении, а также схожесть манифестации суперлативной формы прилагательного и второго лица глагола в среднеанглийском языке в связи с отсутствием ротацизма первой, а

также компаративной формы и личных форм настоящего времени в древнескандинавском в связи с ротацизмом последних. Кроме того, в среднеанглийском языке присутствует значительная вариативность орфографической реализации. В целом же по итогам анализа приходим к выводу, что адекватная векторная репрезентация такого текста, необходимая для обучения алгоритмов частеречной разметки, должна обеспечивать идентичное восприятие двух инстанций одного слова, различающихся написанием на один-два символа, при этом обусловливать различное положение в векторном пространстве морфологически схожих слов, относящихся к разным частям речи. Предполагаем, что такое кодирование может быть реализовано методом посимвольной векторизации с помощью расчета скользящего среднего.

Обзор литературы показал, что в настоящий момент проблематика частеречной разметки исторического текста не изучена до конца и оставляет большой простор для исследования. Можно выделить несколько основных проблем.

Во-первых, такие исследования, как правило, задействуют лишь единичный алгоритм разметки, не прибегая к ансамблированию, т.е. к построению моделей, которые использовали бы сразу несколько базовых алгоритмов и принимали окончательное решение на основании их выходных результатов. Между тем эффективность алгоритма может быть контекстуально зависимой, что говорит об уместности комбинирования разных моделей и методов машинного обучения. Единственной найденной работой, которая задействует комбинирование, является работа И. Яна и Ж. Айзенштайна, в которой, однако, ансамблирование используется лишь для коррекции результатов единичных алгоритмов (опорные векторы и CMM).

Во-вторых, задействованные алгоритмы, показавшие наибольшую эффективность, применяют либо параллелизацию нескольких разных текстов (т.е. обучение на более современном тексте в целях последующего распространения алгоритма на более старый), либо перекрестную проверку с предварительном обучением на всем объеме исторического корпуса или

субкорпуса или значительной его части. При постановке задачи настоящего исследования одним из основных положений является полное отсутствие предварительно подготовленного материала обучающей выборки большого объема, что исключает применение как параллельного текста, так и обучение на всем объеме корпуса: необходимо добиться высокой эффективности на малых выборках, так как конечной целью является упрощение подготовки собственных корпусов малого объема усилиями одного лингвиста или небольшой группы исследователей.

В-третьих, описанные работы преимущественно опираются на применение алгоритмов, функционирование которых основано на полиграммах, т.е. на представлении лексических последовательностей, что не позволяет корректно классифицировать слова вне их контекста. Репрезентация морфологических признаков лексемы практически не задействуется, хотя, на наш взгляд, сочетание именно морфологической презентации и «встраивания» могло бы обеспечить большую точность и гибкость алгоритмизации машинного обучения.

Все это позволяет говорить о новаторском подходе настоящего исследования, в котором планируется апробировать дистрибутивно-независимый метод векторизации текста в сочетании с ансамблевым машинным обучением в целях автоматизации одного из важнейших процессов составления корпуса, обращаясь при этом к малой обучающей выборке.

Тем не менее, результаты, полученные другими авторами, ценны с точки зрения реализации намеченных задач, в частности, с точки зрения проецирования данных, извлеченных из текстов одного периода, на субкорпус другого периода. Это определяет еще одну новаторскую особенность нашей разработки: вовлечение метаданных о тексте в процесс машинного обучения. Применение же обучающих выборок малого размера потенциально позволит вести работу с корпусами тех языков, диахронический материал которых сильно ограничен в объеме и к настоящему времени еще не был размечен вручную, что, на наш взгляд, придает диссертационному исследованию большую практическую значимость, не ограничиваемую областью германистики.

ГЛАВА II. АНСАМБЛЬ-МОДЕЛЬ АВТОМАТИЧЕСКОГО АННОТАТОРА ИСТОРИЧЕСКИХ ТЕКСТОВ И ЕЕ ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА

2.1. Корпусный материал и его подготовка

Для проведения эксперимента по обучению какого-либо алгоритма необходимы текстовые данные, содержащие частеречную разметку. В качестве таковых нами были выбраны Лингвистический атлас раннего среднеанглийского языка [Laing, 2013], далее — LAEME; и Архив средневековых скандинавских текстов [Medieval Nordic Text Archive (Menota)], из которых первый находится в открытом доступе, второй размещен на платформе Clarino, доступ к которой был получен в рамках учебно-исследовательского проекта автора диссертации во время обучения в университете г. Берген, Норвегия, далее – Menota.

В то время как сам по себе LAEME в том виде, в котором он доступен на сайте проекта, не содержит частеречной разметки как таковой, на платформе GitHub выложена полностью размеченная версия (PLAEME, Parsed LAEME, здесь и далее описание корпуса приводится по [Truswell и др., 2019]) в виде набора файлов формата .psd, предназначенных для работы с утилитой CorpusSearch [CorpusSearch | Home]. PLAEME использует тексты из LAEME, датированные 1250–1325 гг., и добавляет синтаксическую разметку того же формата, что задействован в Penn Parsed Corpora of Historical English, далее — PPCHE [Kroch, Taylor, 2000]. Выборка текстов имеет следующие ограничения:

- объем текста — не менее 100 слов (измеряется по числу токенов);
- исключаются тексты, содержащихся в других полностью размеченных корпусах среднеанглийского языка, таких, как Penn-Helsinki Parsed Corpus of Middle English или Parsed Corpus of Middle English Poetry;
- по одной версии каждого текста. В то время как оригинальный LAEME может содержать несколько вариантов, в том числе диалектных, при составлении PLAEME варианты подбирались так, чтобы обеспечить равномерную представленность диалектных вариаций языка, а в случае невозможности выбора

по такому критерию – наиболее объемный вариант. Таким образом, генеральная лингвистическая совокупность PLAEME сокращена до 68 текстовых единиц общим объемом около 172 тысяч слов-токенов.

Ввиду ограниченности кодировок, с которыми может работать утилита CorpusSearch, создатели PLAEME сократили алфавит до стандартной латиницы, избавившись от символов рунического происхождения и акцентированных букв, при этом используются следующие соответствия:

Таблица 2.

Условные обозначения в PLAEME

+a/+A	æ/Æ
+d/+D	ð/Ð
+t/+T	þ/Þ
+g/+G	ȝ/Ȝ
+g2/+G2	ɠ/ɠ
+w/+W	ƿ/Ƿ
‘	Акцент-акут
^	Надстрочный символ

В связи с необходимостью произвести посимвольную векторизацию текста, при которой такие графемы, как торн или эт, могут быть форморазличительными, нами принято решение отказаться от подобных обозначений и привести их к оригинальной форме; акцент «акут» в корпусе встречается редко (менее чем у 0,1% слов), в связи с чем был удален полностью в целях сокращения размерности. Надстрочные символы в конечном векторизуемом материале приведены к обычным строчным. Таким образом, конечный используемый в настоящем исследовании алфавит состоит из 35 символов.

Как LAEME, так и PLAEME составлены с целью сохранить графический характер манускриптов, на основе которых они сформированы, в связи с чем

часть текста в обоих корпусах обособлена: […], что указывает на повреждение манускрипта и невозможность прочтения, а значит, и перевода в электронный вид текста в таких позициях. В отдельных случаях создатели PLAEME производят гипотетическое предположение того, какая графема и/или слово могли находиться в поврежденной и нечитаемой части манускрипта; в других случаях между квадратными скобами оставлено многоточие. В рамках настоящего исследования все элементы с подобной маркировкой удалены из анализируемого материала.

Кроме того, создатели PLAEME используют дополнительную разметку в целях указания исправлений в тексте манускрипта; так, метками >…> обозначены «вставки» – слова, надписанные в манускрипт поверх строк или между другими словами; метками <…< – вычеркнутые писцом слова. Так как в настоящем исследовании не задействуются последовательные контекстозависимые модели, например, основанные на марковских процессах, в рамках настоящего исследования подобные лексемы обрабатываются, как обычные.

Отметим, что текст содержит весьма обширную частеречную разметку, используются 217 помет, при этом многие из них встречаются единично, что может отрицательно сказаться на качестве обучения. Более того, в документации к пакету scikit-learn языка программирования Python [Hackeling, 2014] – основному инструменту настоящего исследования – указывается, что перекрестная проверка не допускает наличие в тестируемой выборке классов, число инстанций которых меньше числа проходов. Количество помет обусловлено тем, что в них отражается составной характер слов, компоненты которых относятся к разным частям речи; в отношении некоторых слов в помете, помимо собственно части речи, указывается субкласс лексемы, например, *Meddeltone* – NPR (noun proper, имя существительное собственное); вспомогательные глаголы выделены в отдельный класс с тернарной оппозицией (BE, DO, HAVE), при этом каждый из них подразделяется далее на несколько подклассов. Отдельная помета используется для модальных глаголов. Кроме

того, как уже было сказано ранее, среднеанглийскому тексту свойственна клитика отрицательной частицы, что находит отражение в разметке LAEME: *nas* = *ne is* = *be not* → NEG+BED. В отдельные одноименные классы выделены местоимение третьего лица man, частица to, многоклассовое слово (прилагательное, наречие или детерминатив) such, детерминатив с адъективальной парадигмой oðer, детерминатив с редуцированной парадигмой on, «один». Помимо прочего, пометы отражают подклассы предлогов и союзов, а также вопросительный характер местоимений и кванторов.

Выглядит разметка в файлах-исходниках корпуса следующим образом:

```
( (IP-MAT (CONJ an-&)
  (NP-SBJ *con*)
  (DOP do=þ-do)
  (PP (NP (PRO me-me))
     (P fro-from))
  (PUNC .)
  (NP-OB1 (CP-FRL (WNP-1 0)
                  (C þat-that)
                  (IP-SUB (NP-OB1 *T*-1)
                          (NP-SBJ (PRO ic-I))
                          (VB [][p]end~-spend)
                          (CODE
{COM_LAEME:first_letter_lost_and_second_partially_lost_because_of_a_wormhole})
                          (NEG-RH n@-ne)
                          (MD @old~-will))))
  (LINEBREAK \))
  (ID ABERDEENT.14))
```

В связи с необходимостью сокращения числа классов и увеличения выборочной лингвистической совокупности в каждом из них в полуавтоматическом режиме с помощью языка регулярных выражений и текстового редактора BBEdit произведена редукция помет по следующим принципам:

- у составных слов, компонентами которых являются прилагательное и существительное, при этом существительное располагается на втором месте и таким образом является главным словом – использовать помету существительного, например, *norþside*, «северная сторона»: ADJ+N → N;

- аналогичным образом поступаем со всеми прочими составными словами, где главным словом является существительное;
- все вспомогательные глаголы всех подклассов сводятся в одну часть речи с пометой AUX;
- модальные глаголы сохраняют свою помету MD;
- вопросительный характер местоимений и кванторов игнорируется;
- слова to, such, one, other, man сохраняют свои уникальные пометы;
- так как отрицательные глаголы, частица которых реализована клитически, как правило, редуцируются в корне (ср. *nele* и *ne wille*), а морфографемическое кодирование обусловливает несколько больший вес символов в начале слова, подобные глаголы нами выделены в отдельный класс NEV с соответствующей пометой;
- к словам, являющимся сочетанием квантора и местоимения, например, somewho, применяется помета квантора;
- в сочетании глагола с любой частью речи, кроме существительного, применяется маркировка глагола;
- в сочетании слова, составляющего собственный класс, с другой частью речи, используется помета по второй лексеме, например, *onepousent*, «одна тысяча» = one thousand = ONE+NUM → NUM;
- все детерминатив-содержащие слова определены как детерминативы, если иное не указано выше.

Логика означенных изменений заключается в том, что в рамках настоящего исследования используется морфографемическое кодирование текста, поэтому окончательное частеречное деление сформировано таким образом, чтобы обеспечить более четкое соответствие помет и морфологических компонентов в ассоциированных с ними лексемах. За счет подобных изменений удалось сократить общее число помет с 217 до 19 (обращаем внимание, что в качестве базиса так или иначе использовались те пометы, которые уже содержатся в корпусе): ADJ (прилагательное), N (существительное), ADV (наречие), AUX (вспомогательный глагол), D (детерминатив), CONJ (союз), NUM

(числительное), PRO (местоимение), INTJ (междометие), MAN (местоимение третьего лица с общим значением «человек»), MD (модальный глагол), NEG (отрицательная частица), NEV (глагол с клитикой отрицательной частицы), NWARD (существительное + наречный суффикс -ward, функционально является наречием и выступает в качестве обстоятельства образа действия – направления движения, однако в соответствии с мотивами маркировки корпуса LAEME также выделяется в отдельный класс), ONE, OTHER, P (предлог), Q (квантор), SUCH, TO (глагольная частица – маркер инфинитива), V (глагол). Распределение частей речи при полученном размечивании показано на рисунке 4.

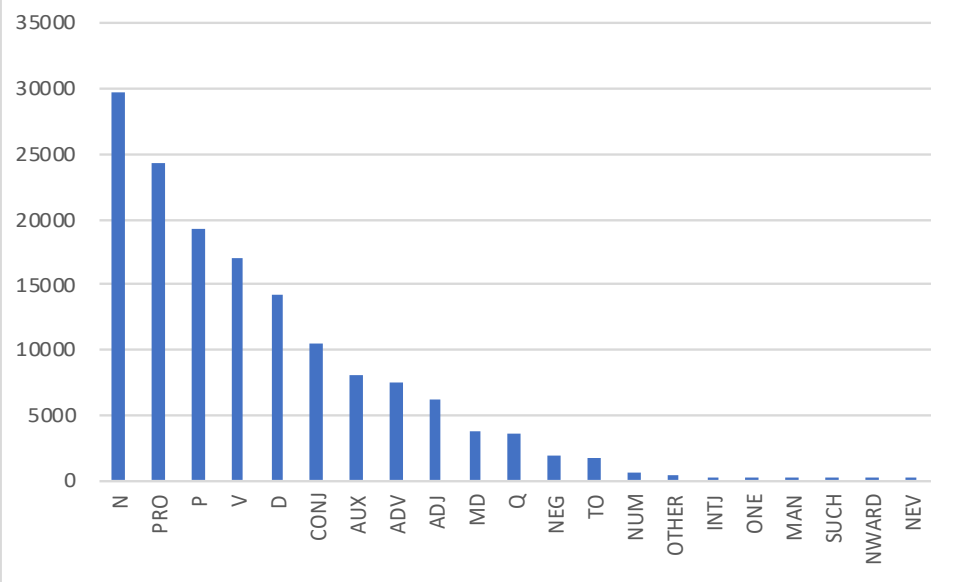


Рис. 7. Распределение частей речи в анализируемых образцах среднеанглийского языка

В рамках экспериментального анализа древнескандинавского языка из корпуса Menota был взят один текст – Konungs skuggsjá, «Царское зерцало», представляющий из себя образовательный трактат, датируемый примерно 1250 года и изначально написанный в целях воспитания короля Магнуса Лагабёте, сына Хокона Хоконссона, в виде диалога между сыном и отцом, где первый задает второму различные вопросы о политике, морали и законе, рыцарском кодексе чести, военной стратегии и тактике, отношениях государства и церкви. Текст сохранился полностью и доступен в составе Menota в аннотированной по частям речи, морфологии и синтаксису форме; источником текста является

манускрипт, кодированный как AM 243 и хранящийся в г. Берген [Haugen, 1994]. Совокупный объем текста составляет примерно 60 тысяч слов.

К сожалению, документация к платформе Clarino [Corpuscle Documentation, 2019] не разъясняет в подробностях значение тех или иных помет, содержащихся в .xml-файлах, которыми наполнен Menota, в связи с чем извлечение непосредственно частеречных помет было произведено по наитию. Ниже приведен пример разметки, содержащейся в исходном файле анализируемого текста:

<s>

<lb ed="ms" n="18"/><w xml:id="w0002" lemma="góðr" me:orig-lemma="góðr" me:msa="xAJ rP gM nS cA sI"><me:dipl><c type="initial"

rend="red blue">G</c>oðan</me:dipl></w>

<w xml:id="w0003" lemma="dagr" me:orig-lemma="dagr" me:msa="xNC gM nS cA sI"><me:dipl>dag</me:dipl></w>

<lb ed="ms" n="19"/><w xml:id="w0004" lemma="herra" me:orig-lemma="herra" me:msa="xNC gM nS cN sI">

<me:dipl>hærra</me:dipl></w>

<w xml:id="w0005" lemma="minn" me:orig-lemma="minn" me:msa="xDP gM nS cN"><me:dipl>min<ex>n</ex></me:dipl></w><pc>.</pc>

</s>

Очевидно, для обозначения морфосинтактической информации в данном формате используется тег me:msa, где первая после открывающей кавычки помета «xXX» обозначает часть речи и подкласс в ее рамках, далее следуют характерные только для данной части речи слова свойства, например, мужской род (gM), единственное число (nS) и винительный падеж (cA). В отличие от PLAEME, разметка данного текста также указывает лемму. Кроме того, судя по тегу <me:dipl>, в исследуемом образце корпусной единицы используется так называемая дипломатическая запись – принцип транскрибирования, в основе которого лежит цель обеспечить легко читаемое, но в то же время точное воспроизведение оригинального манускрипта. В отличие от факсимильной записи, дипломатическая, как правило, раскрывает оригинальные графемические аббревиатуры, например, *v̓a* → *vera*, что упрощает ее понимание,

а также редуцирует палеографические различия, что сокращает орфографическую вариацию и облегчает обработку подобного текста; таким образом, дипломатическая запись использует де-факто нормализованную орфографию с устранением аллографических вариаций, которыми насыщен среднеанглийский корпус. В то время как адекватное сопоставление результатов анализа среднеанглийского и древнескандинавского текстов в рамках машинного обучения при использовании идентичных алгоритмов и метода векторизации требует применения исходного текста со схожими характеристиками, одной из проблем является различие конвенций корпусной лингвистики: Menota сама по себе не содержит достаточно объемных единичных текстов в факсимильной записи, а дипломатическая запись не является широко распространенным способом представления среднеанглийского текста, независимо от корпуса.

Как и в случае со среднеанглийским текстом, оригинальный текст содержит достаточно большое количество частеречных помет с указанием подклассов (порядка 100), при этом отсутствуют пометы, обозначающие комбинированные лексемы, слова в составе которых относились бы к разным частям речи; кроме того, древнескандинавскому языку была несвойственна клитическая реализация отрицательных частиц и личных местоимений, в связи с чем отсутствуют слова, часть речи которых была бы комбинирована естественным образом в связи с подобной графической реализацией. Кроме того, платфома Clarino не использует отдельных частеречных помет, зарезервированных под конкретные лексемы, в связи с чем в применяемой классификации нет классов, одноименных со словами, в них содержащимися. В связи с этим задача редукции набора частеречных помет достаточно проста и сводится к удалению информации о подклассе в рамках той или иной части речи, например. xVB → VERB, xNS → NOUN и т.д.

По итогам выполненной редукции остается всего 11 классов: NOUN (существительное), VERB (глагол), CONJ (союз), DET (детерминатив), ADV (наречие), PREP (предлог), PRON (местоимение), ADJ (прилагательное), PART

(частица), а также исключительно малоразмерные (< 10 примеров на каждый) классы xAT, xAQ, природу которых установить эмпирически не удалось и которые также не поясняются в документации к платформе Clarino. Обратим внимание, что при работе с древнескандинавским языком нами было принято решение отказаться от частеречных помет, используемых в самом тексте корпуса, и вместо них обратиться к лейпцигской грамматической нотации. Отметим также, что в отличие от PLAEME, оригинальные файлы Menota предназначены для работы с поисковой системой Corpuscle, где они размещены; данный инструмент, в отличие от программы CorpusSearch, был разработан относительно недавно и полностью поддерживает кодировку UTF-8 [Yergeau, 2003], в связи с чем создатели корпуса не используют никаких символов-заменителей. Поэтому в оригинальном файле присутствуют эт, торн, акцентированные символы и прочие буквы, не входящие в стандартную 26-значную латиницу, никаких посимвольных замен при подготовке текста к анализу не производится. Совокупный размер алфавита составляет 47 символов.

Распределение частей речи при полученном размечивании выглядит следующим образом:

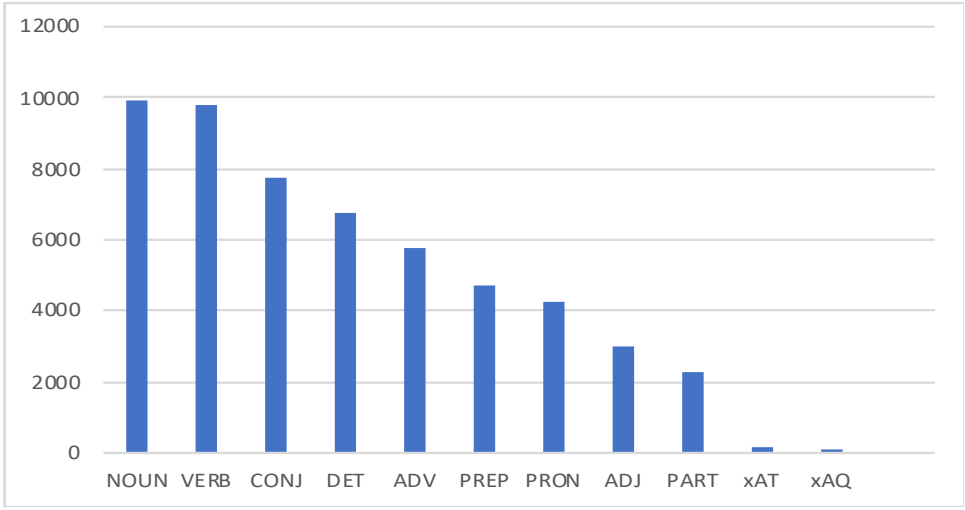


Рис. 8. Распределение частей речи в анализируемом образце древнескандинавского языка

Интересно отметить, что хотя в обоих языках, исходя из статистики по имеющемуся материалу, наиболее распространенной частью речи является имя существительное, распределение прочих несколько отличается, в частности,

местоимения используются значительно реже в древнескандинавском тексте; при этом в обоих языках глагол частотнее наречия, а наречие – прилагательного. С точки зрения обучения частеречного классификатора наибольший интерес представляют морфологически коллидирующие части речи – именные, прилагательное и наречие, прилагательное и глагол – все из которых являются достаточно частотными. В целом распределение по частям речи следует закону Парето, одной из реализаций которого является закон Ципфа [Manning, Schütze, 2000], что, хотя не имеет прямого отношения к проблематике настоящей работы, само по себе является, на наш взгляд, интересным наблюдением.

2.2. Среднеанглийский язык: разметка отдельными алгоритмами и ансамблем

По итогам подготовки среднеанглийского текста к эксперименту получена таблица, состоящая из ~170 тысяч строк текста и 106 столбцов, где первые 105 содержали числовые значения, рассчитанные методом скользящего среднего (35 символов алфавита x3), последний — частеречную помету. Данная таблица была обработана в среде Spyder, оптимизированной для использования языка программирования Python в целях проведения научных расчетов [Доля, 2016], с помощью пакета sklearn, содержащего все перечисленные в первой главе алгоритмы машинного обучения. Результаты проверки приводятся далее отдельно по каждой модели в табличной форме с указанием следующих показателей:

- точность – соотношение $n_{C,t} \in n_C$ / n_C, где n_C – число примеров, отнесенных к классу C, $n_{C,t}$ – число примеров, действительно относящихся к классу C;
- полнота – соотношение $n_C \in n_{C,a}$ / $n_{C,a}$, где $n_{C,a}$ – совокупное число примеров класса C во всей выборке;
- F-показатель – удвоенное значение дроби, числителем которой является произведение точности и полноты, знаменателем – их сумма; один из

традиционных показателей статистического тестирования [Sokolova, Japkowicz, Szpakowicz, 2006]:

$$F = 2\frac{PR}{P+R},$$

где P — точность (англ. precision), R — полнота (англ. recall).

Поскольку одной из целей работы является поиск инструмента ручного формирования корпуса для личных исследовательских целей лингвистов, экспериментальная проверка проводилась на маломощном, «домашнем» вычислительном оборудовании (ЭВМ MacMini 2018: IntelCore i7-8700B, 16 Гб DDR4-памяти 2666 МГц). Среднее время обучения приводимых здесь моделей на полном объеме выборки составляло ~18 минут, что считаем приемлемым для их последующей реализации в рамках корпусного редактора для индивидуального пользования [Каримов, Акинин, 2017] (обратим внимание, что время обучения не равно времени тестирования). Тестирование каждой модели проводилось на той же выборке, которая использовалась в обучении, распространенным методом 10-проходной перекрестной проверки.

Самым простым из представленных в первой главе настоящей работы алгоритмов машинного обучения является метод k ближайших соседей; используя k=5, получили следующие результаты[5]:

Таблица 3.

Результат классификации среднеанглийских лексем по частям речи методом k ближайших соседей

Точность	**Полнота**	**F-показатель**	**Класс**
0.739	0.738	0.739	ADJ
0.892	0.894	0.893	N
0.771	0.794	0.782	ADV
0.886	0.942	0.914	AUX
0.906	0.968	0.936	D
0.977	0.860	0.915	CONJ
0.865	0.806	0.834	NUM
0.969	0.924	0.946	PRO

[5] Все алгоритмы машинного обучения исполнялись в интегрированной среде разработки Spyder языка Python с использованием библиотеки машинного обучения scikit-learn.

0.805	0.560	0.660	INTJ
?[6]	0.000	0.000	MAN
0.870	0.934	0.901	MD
0.857	0.964	0.907	NEG
0.667	0.471	0.552	NEV
0.889	0.842	0.865	NWARD
0.625	0.023	0.044	ONE
0.825	0.976	0.894	OTHER
0.814	0.941	0.873	P
0.915	0.921	0.918	Q
0.919	0.941	0.930	SUCH
?	0.000	?	TO
0.875	0.833	0.854	V
0,883	0.883	0.883	Средневзв. значение

Всего 88,3114% содержащихся в выборке примеров были классифицированы корректно, что превышает результаты большинства цитируемых по тематике работ. В отношении некоторых классов наблюдается низкий показатель точности на уровне ~77%, в связи с чем представляется необходимым проанализировать матрицу ошибок классификатора:

Таблица 4.

Матрица ошибок классификатора kNN

a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**Классиф. как**		
4607	721	395	24	10	0	14	33	10	0	**a**	=	ADJ
660	26496	376	67	8	18	33	139	9	0	**b**	=	N
348	333	6042	68	29	16	8	33	2	0	**c**	=	ADV
13	85	8	7642	3	9	0	220	1	0	**d**	=	AUX
1	14	163	8	13799	0	6	55	0	0	**e**	=	D
1	7	15	1	62	9099	1	4	0	0	**f**	=	CONJ
5	56	6	0	2	0	556	4	0	0	**g**	=	NUM
12	52	170	690	843	1	4	22426	6	0	**h**	=	PRO
7	21	10	1	28	0	0	27	136	0	**i**	=	INTJ
1	95	0	0	0	0	0	97	0	0	**j**	=	MAN
29	138	29	0	3	5	0	1	0	0	**k**	=	MD
0	2	1	0	0	2	0	0	0	0	**l**	=	NEG

[6] Здесь и далее в таблицах классификации знак вопроса означает, что при тестировании ни одному объекту в выборке не был присвоен соответствующий класс.

0	3	1	1	0	0	0	0	0	0	**m**	=	NEV
0	2	2	0	81	1	0	0	0	0	**p**	=	ONE
6	1	0	0	1	1	0	0	0	0	**q**	=	OTHER
27	113	434	77	206	110	3	63	1	0	**r**	=	P
87	48	7	1	1	47	0	4	0	0	**s**	=	Q
1	3	1	0	0	0	0	0	0	0	**t**	=	SUCH
0	0	0	2	2	0	0	3	0	0	**u**	=	TO
428	1521	175	39	161	4	18	35	4	0	**v**	=	V

k	**l**	**m**	**p**	**q**	**r**	**s**	**t**	**u**	**v**	**Классиф. как**		
45	0	0	0	6	40	79	1	0	256	**a**	=	ADJ
183	0	0	0	6	163	50	5	0	1418	**b**	=	N
27	0	1	0	0	596	6	1	0	99	**c**	=	ADV
1	0	1	0	0	93	1	0	0	32	**d**	=	AUX
0	0	0	1	0	159	1	0	0	45	**e**	=	D
3	251	0	0	68	981	84	0	0	1	**f**	=	CONJ
1	0	0	0	0	50	1	0	0	9	**g**	=	NUM
2	0	0	0	0	46	0	0	0	30	**h**	=	PRO
0	0	0	0	0	4	3	0	0	6	**i**	=	INTJ
1	0	0	0	0	0	0	0	0	0	**j**	=	MAN
3586	0	0	0	1	6	1	1	0	40	**k**	=	MD
1	1921	0	0	0	0	62	0	0	3	**l**	=	NEG
1	1	8	0	0	0	0	0	0	2	**m**	=	NEV
0	0	0	5	0	122	0	0	0	4	**p**	=	ONE
0	0	0	0	400	0	0	0	0	1	**q**	=	OTHER
8	1	0	0	0	18066	11	3	0	70	**r**	=	P
3	65	1	0	3	1	3308	0	0	15	**s**	=	Q
1	0	0	0	0	3	0	159	0	1	**t**	=	SUCH
0	0	0	0	0	1682	0	0	0	0	**u**	=	TO
257	3	1	2	1	171	10	3	0	14173	**v**	=	V

Из матрицы ошибок видно, что значительная часть прилагательных отнесена классификатором к существительным, и в меньшей степени – к наречиям, а вот ожидаемая коллизия с глаголом практически отсутствует (см. таблицу 4, часть 2). В равной степени наблюдаем коллизию наречия с прилагательным и существительным. Глагол мало подвержен коллизии с какими-либо частями речи, кроме существительного, что может быть обусловлено редукцией разметки (см. п. «Корпусный материал и его подготовка»).

Отметим, что у служебных частей речи практически отсутствует коллизия между собой, при этом именно у предлогов и союзов наблюдается наиболее высокое значение точности (до 97%). Таким образом, наиболее проблематичной частью речи становится прилагательное, что сходится с результатами, полученными ранее [Karimov, 2018]. Не оправдывает себя выделение местоимения *man* в отдельную часть речи: точность присвоения одноименного класса нулевая, практически все примеры этого местоимения классифицированы как существительные, вероятно, из-за полной омонимии с существительным *man*. Аналогичное явление наблюдается в отношении частицы to, которую данный классификатор определяет как предлог, что, скорее всего, обусловлено наличием в тексте омонимичного предлога; возможно, данную проблему в перспективе можно решить за счет внедрения дистрибутивной репрезентации как дополнительного слоя входных данных, так как предлог и частица *to* различаются синтаксическим распределением. При этом отрицательная частица ne, no, not не выказывает подобных проблем и классифицируется достаточно точно, а ее коллизия с отрицательным глаголом идет в ущерб показателям последнего. Интересен факт коллизии наречия с предлогом, что, возможно, связано с появлением в среднеанглийском языке составных наречий с where и there, таких, как *thereby*.

Еще одним из заявленных ранее алгоритмов является модель случайного леса, 10-проходная перекрестная проверка которой дала следующие результаты:

Таблица 5.

Результат классификации среднеанглийской лексики алгоритмом случайного леса

Точность	**Полнота**	**F-показатель**	**Класс**
0.817	0.723	0.767	ADJ
0.887	0.923	0.905	N
0.800	0.799	0.800	ADV
0.891	0.942	0.916	AUX
0.907	0.968	0.937	D
0.978	0.860	0.915	CONJ
0.913	0.778	0.840	NUM

0.971	0.925	0.947	PRO
0.865	0.556	0.677	INTJ
?	0.000	0.000	MAN
0.870	0.935	0.902	MD
0.855	0.964	0.906	NEG
0.727	0.471	0.571	NEV
1.000	1.000	1.000	NWARD
0.625	0.023	0.044	ONE
0.840	0.971	0.900	OTHER
0.818	0.942	0.876	P
0.924	0.921	0.923	Q
0.982	0.959	0.970	SUCH
?	0.000	0.000	TO
0.893	0.858	0.875	V
0.891	0.891	0.891	Средневзв. значение

Точность классификации составила 89,15%, что превышает результаты kNN, при этом расчетом по критерию Стьюдента [Glickman, Rowntree, 1982] получаем $p \sim 0{,}03$, $p < 0{,}05$[7], что говорит о статистически значимом расхождении при попарном сравнении результатов. Стоит отметить 100% точность и полноту слов класса NWARD; в остальном распределение точности и полноты фактически соответствует аналогичным показателям kNN: высокие значения у служебных слов, детерминативов и имен существительных; несколько пониженные значения у глагола, прилагательного и наречия; высокие значения у вспомогательных и модальных глаголов; неопределение класса у частицы *to* при хороших показателях отрицательной частицы; достаточно точное определение числительных. В целом RFM справился с задачей классификации лучше, хотя по отдельным показателям kNN незначительно превосходит его. Обратимся к анализу матрицы ошибок.

Таблица 6.

Матрица ошибок классификатора RFM на материале среднеанглийского языка

a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	Классифицировано как

[7] Рассчитан в интегрированной среде разработки Spyder языка Python с помощью пакета SciPy.

4513	874	387	16	7	0	8	30	6	0	**a**	=	ADJ
360	27342	258	53	7	16	17	114	4	0	**b**	=	N
291	391	6083	64	29	15	2	27	2	0	**c**	=	ADV
6	90	0	7642	2	8	0	224	1	0	**d**	=	AUX
1	14	163	8	13800	0	6	51	0	0	**e**	=	D
3	13	11	0	62	9097	0	4	0	0	**f**	=	CONJ
2	69	2	0	0	0	537	5	0	0	**g**	=	NUM
7	58	163	685	832	1	3	22453	6	0	**h**	=	PRO
4	28	9	1	28	0	0	26	135	0	**i**	=	INTJ
1	95	0	0	0	0	0	97	0	0	**j**	=	MAN
15	135	25	2	2	5	0	2	0	0	**k**	=	MD
0	1	1	0	0	1	0	0	0	0	**l**	=	NEG
0	5	0	1	0	0	0	0	0	0	**m**	=	NEV
0	0	0	0	0	0	0	0	0	0	**o**	=	NWARD
0	1	1	0	81	1	1	0	0	0	**p**	=	ONE
5	4	0	0	0	1	0	0	0	0	**q**	=	OTHER
7	135	414	75	205	111	2	61	0	0	**r**	=	P
79	65	2	2	1	44	0	6	0	0	**s**	=	Q
0	4	0	0	0	0	0	0	0	0	**t**	=	SUCH
0	0	0	2	2	0	0	3	0	0	**u**	=	TO
229	1484	85	29	159	3	12	19	2	0	**v**	=	V

k	**l**	**m**	**o**	**p**	**q**	**r**	**s**	**t**	**u**	**v**	Классифицировано как		
47	0	0	0	0	4	24	78	0	0	247	**a**	=	ADJ
187	1	0	0	0	0	124	19	0	0	1130	**b**	=	N
26	1	1	0	0	0	589	4	0	0	84	**c**	=	ADV
1	0	1	0	0	0	94	1	0	0	39	**d**	=	AUX
0	0	0	0	1	0	158	1	0	0	49	**e**	=	D
4	251	0	0	0	67	980	86	0	0	0	**f**	=	CONJ
1	0	0	0	0	0	51	0	0	0	23	**g**	=	NUM
0	0	0	0	0	1	56	0	0	0	17	**h**	=	PRO
0	0	0	0	0	0	4	3	0	0	5	**i**	=	INTJ
1	0	0	0	0	0	0	0	0	0	0	**j**	=	MAN
3592	0	0	0	0	1	8	0	0	0	53	**k**	=	MD
1	1921	0	0	0	0	0	63	0	0	4	**l**	=	NEG
1	1	8	0	0	0	0	0	0	0	1	**m**	=	NEV
0	0	0	19	0	0	0	0	0	0	0	**o**	=	NWARD
0	0	0	0	5	0	122	1	0	0	4	**p**	=	ONE
1	0	0	0	0	398	0	0	0	0	1	**q**	=	OTHER
11	1	0	0	0	0	18088	9	3	0	71	**r**	=	P
2	69	0	0	0	3	2	3307	0	0	9	**s**	=	Q
0	0	0	0	0	0	0	0	162	0	3	**t**	=	SUCH
0	0	0	0	0	0	1682	0	0	0	0	**u**	=	TO
252	2	1	0	2	0	130	6	0	0	14591	**v**	=	V

Здесь наблюдаем схожие с методом ближайших соседей паттерны: прилагательное оказывается в значительной коллизии с существительным, в меньшей степени с наречием, при этом полностью отсутствует его коллизия с глаголом; с другой стороны, сам глагол демонстрирует некоторую коллизию с прилагательными, хотя в меньшей мере, чем с существительными. Вспомогательный глагол, отрицательная частица, модальный глагол практически ни с чем не коллидируют, а наречия, отнесенные к классу NWARD, определяются со 100% точностью и полнотой. Низка коллизия союзов, предлогов и кванторов как между собой, так и с самостоятельными частями речи, а вот местоимения в значительной мере пересекаются с другими именными частями речи, с наречием, с детерминативом, что, с одной стороны, может быть обусловлено схожестью, если не идентичностью парадигмы, с другой – омографией некоторых детерминативов и указательных местоимений, ср. *þat cynig* (определенный артикль/указательное местоимение – детерминатив) и *cynig þat* (относительное местоимение). Слова, выделенные в собственные одноименные классы, определяются практически однозначно, а частица *to* не определяется вовсе, вместо этого в 99% случаев отнесена к предлогам, что, скорее всего, обусловлено ее графической идентичностью омонимичному предлогу и для разрешения требует контекстозависимого кодирования.

Таким образом, использование модели случайного леса (в данном случае состояла из 100 деревьев принятия решений) обеспечивает статистически значимое улучшение показателей точности, что видно на следующем рисунке, но фактически сохраняет те же закономерности ошибок классификации.

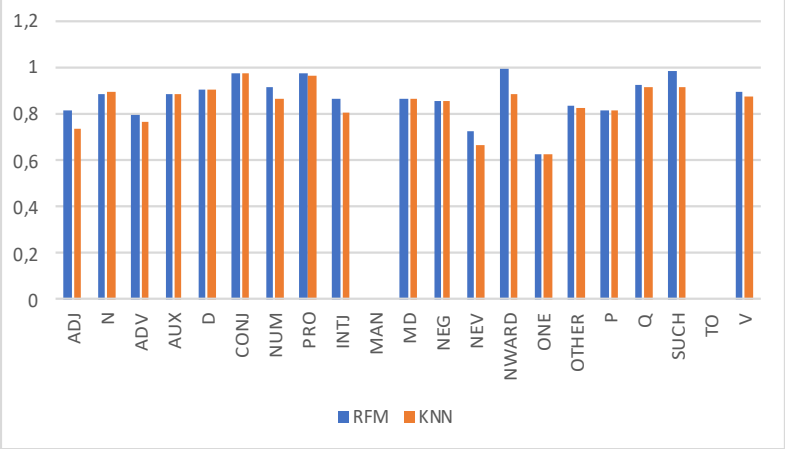


Рис. 9. Сравнение точности классификаторов RFM, KNN по различным частям речи на материале среднеанглийского языка

Остальные из заявленных ранее алгоритмов – метод опорных векторов, наивный байесовский классификатор, логистическая регрессия и многослойный перцептрон – не продемонстрировали выдающихся результатов, что противоречит результатам, полученным в отношении среднеанглийского языка ранее [Karimov, Akinin, Yakymets, 2018]; данное противоречие не затрагивает метод Naïve Bayes, который и ранее был определен как наименее точный из анализируемых, но в значительной мере касается многослойного перцептрона и метода опорных векторов, точность которых понизилась статистически значимо, что может быть обусловлено характером выборки. В связи с низкой точностью матрицы ошибок по перечисленным алгоритмам не анализируются и не приводятся.

Точность классификации на основе теоремы Байеса составляет всего 0,37%; нужно отметить, что сама по себе теорема постулирует вероятностный подход и, возможно, требует иного метода векторизации – на уровне слов-токенов и их контекста, а не по символам.

Таблица 7.

Результат классификации среднеанглийской лексики наивным Байесовским алгоритмом

Точность	**Полнота**	**F-показатель**	**Класс**
0.00	0.00	0.00	ADJ
0.06	0.01	0.01	ADV
0.25	0.21	0.23	AUX
0.53	0.89	0.66	CONJ
0.67	0.98	0.80	D
0.00	0.00	0.00	INTJ
0.00	0.00	0.00	MAN
0.00	0.00	0.00	MD
0.49	0.25	0.33	N
0.46	0.56	0.50	NEG
0.00	0.00	0.00	NEV
0.00	0.00	0.00	NUM
0.00	0.00	0.00	NWARD

0.00	0.00	0.00	ONE
0.00	0.00	0.00	OTHER
0.35	0.51	0.41	P
0.55	0.79	0.65	PRO
0.14	0.20	0.17	Q
0.00	0.00	0.00	SUCH
0.00	0.00	0.00	TO
0.16	0.15	0.15	V
0.37	0.44	0.39	Средневзв. значение

Результаты классификации по Байесу в целом характеризуются теми же тенденциями, что и выводы двух ранее рассмотренных классификационных алгоритмов: наиболее высокой точностью отличаются существительное и некоторые служебные части речи, при этом даже их показатели слишком низки, чтобы можно было говорить о каком-либо практическим применении данного классификатора с реализуемым в рамках настоящей работы методом векторизации текста. Кроме того, те части речи, классификация которых не выполняется моделями случайного леса и k ближайших соседей, также не распознаются наивным байесовским алгоритмом, что исключает его дальнейшее внедрение в ансамбль.

Результаты применения метода опорных векторов на порядок лучше, но все равно в значительной мере уступают точности случайного леса и k ближайших соседей:

Таблица 8.

Результат классификации среднеанглийской лексики методом опорных векторов

Точность	**Полнота**	**F-показатель**	**Класс**
0.077	0.000	0.000	ADJ
0.445	0.881	0.592	N
0.556	0.366	0.441	ADV
0.699	0.643	0.670	AUX
0.882	0.937	0.909	D
0.934	0.841	0.885	CONJ
?	0.000	?	NUM
0.845	0.878	0.861	PRO
?	0.000	?	INTJ

?	0.000	?	MAN
0.711	0.404	0.515	MD
0.703	0.881	0.782	NEG
?	0.000	?	NEV
?	0.000	?	NWARD
?	0.000	?	ONE
0.000	0.000	0.000	OTHER
0.701	0.687	0.694	P
0.795	0.393	0.526	Q
?	0.000	?	SUCH
?	0.000	?	TO
0.460	0.101	0.165	V
0.649	0.649	0.649	Средневзв. значение

Всего у четырех классов из 20 отмечается точность, сопоставимая с результатами двух наиболее эффективных алгоритмов — союзов, кванторов, местоимений и детерминативов, при этом ни по одному из этих классов метод опорных векторов не обеспечивает более высокую точность и полноту по сравнению со случайным лесом или ближайшими соседями. Хотя по отдельным частям речи удается получить значения, близкие к 90%, общий показатель точности составляет всего 64%. Как уже было сказано ранее, подобная выкладка явно противоречит результатам, опубликованным в рамках текущего исследования ранее [Karimov, 2018], где эффективность данного алгоритма была сопоставима с RFM и превышала показатели kNN. На наш взгляд, подобное расхождение обусловлено тем, что описанный [там же] эксперимент проводился на малой выборке с бинарной классификацией: прилагательные и глаголы; SVM сами по себе являются бинарным классификатором, и при необходимости обработать данные, подразделяющиеся на большее число классов, алгоритм генерирует множество бинарных классификаторов, поочередно противопоставляя каждый класс в выборке всем остальным, чтобы обеспечить дихотомию. Шкалирование данного метода неэффективно ни с точки зрения конечного результата, ни с позиций вычислительного времени; так, обучение модели на каждый проход заняло 389,8 секунды, или около 6 минут, при

использовании наименее ресурсоемкой функции поправки ядра (kernel trick) – линейной.

В предыдущих работах, наряду с методом случайного леса, в качестве наиболее эффективного алгоритма показал себя многослойный перцептрон; данное наблюдение не подтвердилось на выборке, использовавшейся в настоящем исследовании (см. таблицу 8). Точность классификатора составила всего 74,63%, что значительно ниже 87%, полученных ранее. По сравнению с RFM, kNN перцептрон демонстрирует несколько иные паттерны распределения точности по классам; так, существительное уступает любой служебной части речи, в то время как сравнение самостоятельных частей речи между собой по ключевым показателям эффективности классификатора демонстрирует их идентичное отношение: N > V > ADV > ADJ. Отметим также, что в отличие от двух лидирующих классификаторов, многослойный перцептрон не идентифицирует слова класса NWARD, однако относительно успешно (на уровне полноты 97,6%) классифицирует частицу *to*:

Таблица 9.

Результат классификации среднеанглийской лексики многослойным перцептроном

Точность	**Полнота**	**F-показатель**	**Класс**
0.444	0.217	0.291	ADJ
0.610	0.777	0.683	N
0.715	0.524	0.605	ADV
0.849	0.869	0.859	AUX
0.908	0.948	0.927	D
0.968	0.809	0.881	CONJ
0.786	0.096	0.171	NUM
0.920	0.892	0.906	PRO
0.206	0.132	0.161	INTJ
0.000	0.000	0.000	MAN
0.776	0.764	0.770	MD
0.833	0.956	0.890	NEG
?	0.000	?	NEV
0.000	0.000	0.000	NWARD
?	0.000	?	ONE
0.756	0.883	0.814	OTHER

0.819	0.747	0.781	P
0.860	0.656	0.744	Q
0.464	0.077	0.132	SUCH
0.471	0.976	0.636	TO
0.525	0.538	0.532	V
0,746	0.746	0,746	Средневзв. значение

Отметим, что многослойный перцептрон является наиболее ресурсоемким из обсуждаемых алгоритмов; время построения модели составило 1722 секунды, или около получаса, на используемом в рамках эксперимента вычислительном оборудовании, в связи с чем реализация подобного алгоритма в рамках ПО для индивидуальной работы с корпусом не может рассматриваться как практическое решение, ввиду чего в рамках настоящего исследования было решено не включать MLP в состав ансамбля.

Последний из заявленных алгоритмов классификации – функция логистической регрессии – обеспечивает точность выше, чем наивный байесовский классификатор, но ниже, чем любой другой из рассматриваемых методов (см. табл. 10), значительно уступая ведущим алгоритмам: kNN и RFM.

Таблица 10.

Результат классификации среднеанглийской лексики функцией логистической регрессии

Точность	**Полнота**	**F-показатель**	**Класс**
0.00	0.00	0.00	ADJ
0.08	0.00	0.01	ADV
0.38	0.17	0.23	AUX
0.80	0.85	0.82	CONJ
0.75	0.95	0.84	D
0.00	0.00	0.00	INTJ
0.00	0.00	0.00	MAN
0.51	0.23	0.32	MD
0.47	0.94	0.63	N
0.72	0.76	0.74	NEG
0.00	0.00	0.00	NEV
0.00	0.00	0.00	NUM
0.00	0.00	0.00	NWARD
0.00	0.00	0.00	ONE
0.00	0.00	0.00	OTHER

0.66	0.71	0.68	P
0.71	0.83	0.77	PRO
0.64	0.26	0.37	Q
0.00	0.00	0.00	SUCH
0.00	0.00	0.00	TO
0.53	0.09	0.15	V
0.54	0.60	0.53	Средневзв. значение

Точность определения класса методом ЛР составляет всего 54%, при этом данный классификатор, в отличие от других, лучше прочих самостоятельных частей речи классифицирует глаголы, хотя тенденция к более высоким показателям у служебных частей речи (к которым для удобства относим детерминативы) повторяется и в этом случае.

Таким образом, из всех рассмотренных классификаторов только два отличаются достаточно высокой точностью для включения в ансамбль, при этом kNN не подходит для использования в AdaBoost, так как принцип работы последнего предполагает присвоение входным данным различных весовых коэффициентов при повторном обучении (усилении, «бустинге») базового алгоритма; в scikit-learn доступно несколько реализаций метода ближайших соседей, но ни одна из них не поддерживает аргумент sample_weight, необходимый для присвоения значений «веса», а значит, и для работы ABC. Алгоритм случайного леса сам по себе является ансамблем на основе бэггинга, при этом поддерживает параметр sample_weight, что позволяет производить его усиление. Несмотря на это, «бустинг» случайного леса, как правило, не применяется на практике; зачастую AdaBoost и Random Forest сравниваются в противопоставлении друг с другом как ансамбли, использующие диаметрально противоположные методы корректировки результатов базового классификатора (как и RFM, ABC в наиболее распространенных моделях использует в качестве такого классификатора малоразмерное дерево принятия решений; отличие заключается в том, что первый исполняет базовые алгоритмы параллельно, второй – последовательно) [Wyner и др., 2017]. Тем не менее, нам удалось найти

примеры такого ансамблирования: [Leshem, 2005], в связи с чем было принято решение предпринять попытку воспользоваться им.

Таблица 11.

Результат классификации среднеанглийской лексики методом AdaBooster

Точность	**Полнота**	**F-показатель**	**Класс**
0.814	0.709	0.758	ADJ
0.878	0.925	0.901	N
0.799	0.799	0.799	ADV
0.928	0.896	0.911	AUX
0.912	0.962	0.937	D
0.979	0.860	0.916	CONJ
0.904	0.788	0.842	NUM
0.956	0.940	0.948	PRO
0.864	0.547	0.670	INTJ
?	0.000	?	MAN
0.874	0.930	0.901	MD
0.855	0.963	0.906	NEG
0.800	0.471	0.593	NEV
1.000	1.000	1.000	NWARD
0.286	0.046	0.079	ONE
0.837	0.966	0.897	OTHER
0.820	0.941	0.877	P
0.923	0.921	0.922	Q
0.964	0.947	0.955	SUCH
?	0.000	?	TO
0.887	0.850	0.869	V
0.8895	0.890	0.890	Средневзв. значение

Точность данного варианта составила почти 89%, что, с одной стороны, говорит об улучшении результатов по сравнению с «неусиленным», случайным лесом; с другой – разница не представляется значительной. Кроме того, проверка по критерию Стьюдента [Glickman, Rowntree, 1982] возвращает значение p = 0,4303, указывающее на отсутствие статистически значимой разницы между результатами AdaBooster и RFM. Таким образом, два наиболее эффективных алгоритма классификации являются ансамбль-методами; несмотря на то, что бустинг обеспечивает некоторое улучшение результатов моделирования

методом случайного леса, оно статистически недостоверно, в связи с чем применение данного похода на практике считаем неоправданным ввиду значительной вычислительной емкости алгоритма.

Ансамблирование методом стекинга позволяет комбинировать несколько алгоритмов с помощью мета-классификатора, обучаемого на выводах базовых моделей; мета-классификатор присваивает класс, сообщаемый тем классификатором, который ранее в процессе подготовки ансамбль-модели продемонстрировал наибольшую точность в отношении аналогичных или схожих примеров. Таким образом, алгоритмы в стеке производят коррекцию ошибок друг друга. Данная процедура эффективна в том случае, когда базовые классификаторы различаются в точности как в лучшую, так и в худшую сторону по отношению к разным классам, что, в отличие от результатов, полученных ранее и описанных в работе [Karimov, Akinin, Yakymets, 2018], не относится к сложившейся ситуации: случайный лес на имеющихся данных по всем классам, кроме имени существительного, демонстрирует более высокую точность по сравнению с методом k случайных соседей, а по имени существительному разница между этими двумя алгоритмами составляет сотые доли процента; добавление многослойного перцептрона – единственного из проанализированных алгоритмов, имеющего положительную точность и полноту классификации частицы *to* – к ансамблю, на наш взгляд, не имеет смысла в виду его высокой ресурсоемкости. В силу всего вышесказанного от использования стекинг-основанного ансамбля в рамках анализа среднеанглийского текста решено отказаться.

2.3. Древнескандинавский язык: разметка отдельными алгоритмами и ансамблем

По итогам подготовки древнескандинавского языкового материала получена таблица, состоящая из 142 столбцов (47 символов алфавита х 3 и один столбец для обозначения класса) и ~52 тысяч строк; несмотря на бóльшую

размерность данной выборки, за счет наличия в ней меньшего числа примеров понизилась вычислительная емкость расчетов, что расширяет применимость некоторых алгоритмов, в частности, AdaBooster, и упрощает задачу ансамблирования на базе потребительского аппаратного обеспечения.

Экспериментальная проверка алгоритмов проводилась в том же порядке и теми же методами, что и в отношении среднеанглийского языка, и позволила получить в целом сходные результаты.

Таблица 12.

Результат классификации древнескандинавской лексики методом k ближайших соседей

Точность	**Полнота**	**F-показатель**	**Класс**
0.933	0.933	0.933	NOUN
0.991	0.993	0.992	DET
0.994	0.995	0.995	PRON
0.949	0.948	0.948	VERB
0.973	0.981	0.977	ADV
0.997	0.998	0.997	PREP
0.998	0.999	0.999	CONJ
1.000	1.000	1.000	PART
0.871	0.852	0.861	ADJ
0.977	0.992	0.984	xAT
0.960	0.990	0.975	xAQ
0.966	0.967	0.967	Средневзв. значение

Таблица 13.

Матрица ошибок классификатора kNN на материале древнескандинавского языка.

a	b	c	d	e	f	g	h	i	j	k	Классиф. как
8942	27	5	318	61	7	0	0	221	2	1	a=NOUN
20	6551	1	10	3	0	7	0	4	0	2	b=DET
5	2	4069	7	3	0	0	0	1	0	1	c=PRON
322	7	9	8979	38	5	4	0	108	0	0	d=VERB
43	7	1	14	5448	0	2	0	37	1	0	e=ADV
2	1	0	6	1	4560	0	0	0	0	0	f=PREP
2	2	4	1	0	0	7581	0	0	0	0	g=CONJ
0	0	0	0	0	0	0	2160	0	0	0	h=PART
250	11	2	127	43	2	0	0	2501	0	0	i=ADJ
0	1	0	0	0	0	0	0	0	125	0	j=xAT

0	0	1	0	0	0	0	0	0	0	96	k=xAQ

Точность классификации составила 96,6597%, что фактически превышает значения, полученные ранее в работах с применением корпуса древнескандинавского языка. Примечательно, что ни одна из частей речи не демонстрирует существенно низких значений точности и полноты; наименьшее значение обоих показателей отмечается у прилагательного, в связи с чем представляется важным проанализировать матрицу ошибок классификатора, из которой видно, что прилагательным свойственна коллизия с именем существительным, с глаголом, в меньшей степени с наречием, причем, как следует из анализа языкового материала, приведенного в п. 1.3 настоящей работы, подобная коллизия ожидаема с точки зрения как характера кодирования, так и орфографических и морфологических особенностей исследуемого языка, в связи с чем понижение точности до показателя 87%, на наш взгляд, можно считать приемлемым результатом. Бóльшая коллизия с глаголом, нежели с наречием, вызывает определенный интерес, но может быть обусловлена тем, что в отличие от среднеанглийского языка, в древнескандинавском наречный суффикс *-leg* был практически уникален для этой части речи и относительно нечасто встречался у прилагательных, а маркеры прилагательных, в свою очередь, не встречались у наречия, хотя коренная составляющая нередко могла совпадать.

Таблица 14.

Результат классификации древнескандинавской лексики алгоритмом случайного леса

Точность	**Полнота**	**F-показатель**	**Класс**
0.918	0.967	0.942	NOUN
0.994	0.992	0.993	DET
1.000	0.996	0.998	PRON
0.959	0.957	0.958	VERB
0.992	0.981	0.986	ADV
1.000	0.997	0.998	PREP
0.999	0.999	0.999	CONJ
1.000	1.000	1.000	PART

0.940	0.819	0.875	ADJ
0.984	1.000	0.992	xAT
1.000	0.990	0.995	xAQ
0.972	0.972	0.972	Средневзв. значение

Из таблицы 14 видно, что алгоритм случайного леса обеспечивает еще более высокую точность на уровне 97,23%, при этом отсутствуют части речи, в классификации которых допускалось бы существенное количество ошибок. Наименьшую точность в данном случае продемонстрировало имя существительное, а наименьшую полноту — имя прилагательное, что побуждает ознакомиться с матрицей ошибок классификатора.

Интересно, что ожидаемой коллизии существительного и прилагательного здесь практически не возникло (хотя прилагательное в основном классифицируется неверно именно как существительное). Вместо этого ошибка классификации имен существительных в основном заключается в присвоении им класса VERB – глагола, что само по себе вызывает определенный интерес, поскольку глагол мало совпадает с именем существительным морфологически, за исключением глагольных окончаний на -r — литеры, которая также служит маркером множественного числа существительных и именительного падежа у мужского рода, возникнув в результате ротацизма протогерманского /z/. Данный феномен, на наш взгляд, требует отдельного исследования.

Таблица 15.

Матрица ошибок классификатора RFM на материале древнескандинавского языка

a	b	c	d	e	f	g	h	i	j	k	Классиф. как
9265	14	1	224	7	0	3	0	70	0	0	a=NOUN
31	6543	0	12	4	0	0	0	4	2	0	b=DET
8	1	4071	3	2	0	2	0	1	0	0	c=PRON
335	11	0	9062	10	2	2	0	50	0	0	d=VERB
54	7	1	16	5445	0	1	0	29	0	0	e=ADV
9	0	0	3	3	4555	0	0	0	0	0	f=PREP
2	0	0	4	1	0	7583	0	0	0	0	g=CONJ
0	0	0	0	0	0	0	2160	0	0	0	h=PART
384	5	0	126	17	0	0	0	2404	0	0	i=ADJ
0	0	0	0	0	0	0	0	0	126	0	j=xAT
0	0	0	1	0	0	0	0	0	0	96	k=xAQ

Имя прилагательное демонстрирует те же паттерны коллизии, что и в случае с kNN: наиболее часто оно воспринимается как существительное, реже как глагол (см. обсуждение возможных причин выше). Редкость коллизии с глаголом является неожиданным открытием в отношении обоих языков, так как и в первом, и во втором на этапе предварительного анализа была выявлена морфологическая схожесть этих частей речи, хотя и обусловленная разными язык-специфичными факторами. В итоге имеем тот же результат, что и на материале среднеанглийского языка: использование модели случайного леса (в данном случае состояла из 100 деревьев принятия решений) обеспечивает статистически значимое улучшение показателей точности, что видно на следующем рисунке, но фактически сохраняет те же закономерности ошибок классификации. Интересно, на наш взгляд, то, что оба классификатора со 100% или близкой к 100% точностью определяют классы некоторых служебных слов: частиц, предлогов и союзов, что, возможно, связано с отсутствием омонимии между ними в древнескандинавском языке, по крайней мере, в использованном для анализа тексте.

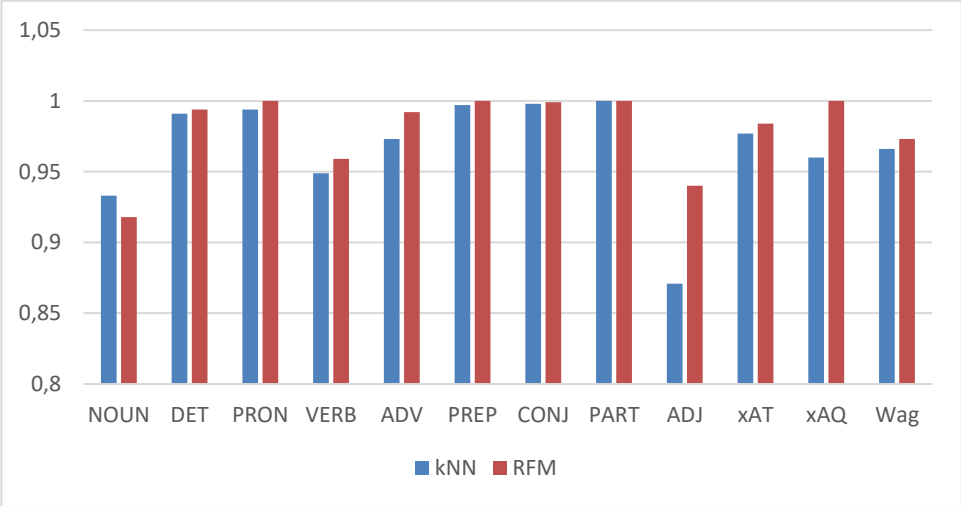


Рис. 10. Сравнение точности классификаторов RFM, KNN по различным частям речи на материале древнескандинавского языка

Если рассматривать прочие классификаторы, то, как и в случае со среднеанглийским языком, они не продемонстрировали высокой точности, хотя их показатели применительно к древнескандинавскому оказались на порядок лучше; при этом ранжирование по качеству классификации фактически следует

тому же порядку – MLP > SVM > LR > NB – что, на наш взгляд, говорит о качественной схожести полученных для двух рассматриваемых языков векторизованных выборок:

Таблица 16.

Результат классификации древнескандинавской лексики наивным байесовским алгоритмом

Точность	**Полнота**	**F-показатель**	**Класс**
0.632	0.221	0.328	NOUN
0.367	0.578	0.449	DET
0.461	0.759	0.573	PRON
0.599	0.203	0.303	VERB
0.458	0.407	0.431	ADV
0.518	0.926	0.664	PREP
0.716	0.986	0.830	CONJ
1.000	1.000	1.000	PART
0.242	0.231	0.236	ADJ
0.642	0.984	0.777	xAT
0.277	1.000	0.434	xAQ
0.556	0.530	0.493	Средневзв. значение

Таблица 17.

Результат классификации древнескандинавской лексики методом опорных векторов

Точность	**Полнота**	**F-показатель**	**Класс**
0.479	0.756	0.587	NOUN
0.802	0.710	0.753	DET
0.879	0.901	0.890	PRON
0.638	0.598	0.617	VERB
0.832	0.668	0.741	ADV
0.930	0.909	0.920	PREP
0.992	0.977	0.984	CONJ
0.999	1.000	0.999	PART
?	0.000	?	ADJ
0.675	0.664	0.669	xAT
?	0.000	?	xAQ
0.735	0.735	0.735	Средневзв. значение

Таблица 18.

Результат классификации древнескандинавской лексики многослойным перцептроном

Точность	**Полнота**	**F-показатель**	**Класс**
0.658	0.814	0.728	NOUN
0.925	0.938	0.932	DET
0.970	0.920	0.944	PRON
0.769	0.784	0.776	VERB
0.937	0.774	0.848	ADV
0.967	0.869	0.915	PREP
0.983	0.966	0.975	CONJ
0.964	1.000	0.982	PART
0.608	0.415	0.493	ADJ
0.692	0.914	0.788	xAT
0.372	0.455	0.409	xAQ
0.848	0.840	0.840	Средневзв. значение

Таблица 19.

Результат классификации древнескандинавской лексики функцией логистической регрессии

Точность	**Полнота**	**F-показатель**	**Класс**
0.72743	0.25437	0.37693	NOUN
0.42241	0.66527	0.51673	DET
0.53061	0.87361	0.66022	PRON
0.68945	0.23365	0.34902	VERB
0.52716	0.46846	0.49608	ADV
0.59622	0.997	0.74620	PREP
0.82412	0.984	0.89670	CONJ
1	1	1	PART
0.27854	0.26588	0.27206	ADJ
0.73894	0.985	0.84441	xAT
0.31883	1	0.48350	xAQ
0.60488	0.70248	0.65004	Средневзв. значение

Стоит отметить, что в то время как метод опорных векторов позволил получить достаточно неплохой на фоне прочих классификаторов результат, некоторые классы при перекрестной проверке не были присвоены ни одному из имеющихся примеров. Наибольшей точностью и полнотой отличился

многослойный перцептрон, тем не менее, уступив как методу *k* ближайших соседей, так и случайным лесам; в связи с этим ни один из четырех дополнительных алгоритмов не может использоваться в составе ансамбля. В итоге ансамблирование решено провести усилением результата, полученного по алгоритму RFM, с помощью AdaBooster: подобный подход ранее обеспечил статистически значимое улучшение точности на материале среднеанглийского языка. При этом получили следующий результат:

Таблица 20.

Результат классификации древнескандинавской лексики методом AdaBooster

Точность	**Полнота**	**F-показатель**	**Класс**
0.921	0.967	0.943	NOUN
0.995	0.994	0.994	DET
0.999	0.996	0.998	PRON
0.961	0.959	0.960	VERB
0.991	0.983	0.987	ADV
0.999	0.997	0.998	PREP
0.999	0.999	0.999	CONJ
1.000	1.000	1.000	PART
0.940	0.818	0.875	ADJ
1.000	1.000	1.000	xAT
1.000	0.990	0.995	xAQ
0.973	0.973	0.973	Средневзв. значение

Точность классификации составила 97,32%, или примерно на 0,9% выше, чем у неусиленного случайного леса; при этом AdaBooster повысил точность классификации существительных – основного «слабого места» RFM, хотя и не обеспечил той же точности их классификации, которая наблюдалась с применением kNN. Тем не менее, разница между двумя алгоритмами статистически недостоверна: $p = 0{,}2678$ по критерию Стьюдента; однако при этом вычислительная емкость расчета AdaBooster на данной выборке относительно невысока, в связи с чем применение подобного метода ансамблирования может быть целесообразно.

Ансамблирование методом стекинга в данном случае, на наш взгляд, не имеет смысла; несмотря на то, что kNN обеспечивает несколько более высокую точность определения существительных, достигается такой результат в ущерб полноте – не менее важному показателю. Таким образом, именно случайный лес/бустинг является, по нашему мнению, рекомендованным методом классификации применительно к исследованному материалу вкупе с использованным подходом к векторизации.

2.4. Сравнительный анализ проявления диахронических различий в контексте классификации: обсуждение результатов эксперимента

По итогам эксперимента установлены следующие ключевые особенности применения алгоритмов машинного обучения к имеющимся данным с векторизацией по методу скользящего среднего:

Наиболее точными оказались алгоритмы *k* ближайших соседей и случайный лес, уровень точности и полноты обоих составил порядка 88% в отношении среднеанглийского языка, 96% и 97%, соответственно, применительно к древнескандинавскому; второй в целом работает точнее, при этом разница между ними статистически значима, что говорит о предпочтительности применения модели RFM в поставленной задаче. При этом оба алгоритма допускают схожие ошибки: так, оба не справляются с присвоением корректного класса глагольной частицы *to* и местоимения *man*, что, на наш взгляд, обусловлено их омонимией (в том числе омографией) с другими частями речи: предлогом *to* и существительным *man(n)*, соответственно. Данное предположение подтверждается анализом матриц ошибок алгоритмов, где указано, к каким классам были некорректно отнесены данные лексемы. Интересно, что древнескандинавскому языку в меньшей мере свойственны ошибки классификации служебных слов, что, как уже было сказано ранее, на наш взгляд, связано с их меньшей омонимией.

Оба алгоритма демонстрируют схожие паттерны ошибок: коллизию прилагательного с существительным и наречием, но не с глаголом, что отличается от ситуации с древнескандинавским языком, где имеется коллизия прилагательного с существительным и глаголом, в гораздо меньшей степени – с наречием. Именно прилагательное демонстрирует наименьшие значения точности и полноты среди всех самостоятельных частей речи, которым в обоих языках свойственны худшие показатели по сравнению со служебными словами. Глагол ни в том, ни в другом языке не проявляет значительной коллизии с другими частями речи, являясь наиболее точно и полно классифицируемой категорией среди неслужебных слов.

Эти и другие особенности рассматриваемых текстов, а также их проявления в рамках проведенных экспериментов, сведены далее в таблице 21.

Таблица 21.

Сравнительный анализ проявлений языковой специфики в результатах вычислительного эксперимента с машинным обучением

Аспект	**Специфика в среднеангл.**	**Эксперимент. проявление**	**Специфика в древнесканд.**	**Эксперимент. проявление**
Сущ.	Использует 4-падежную парадигму. Морфологически близко к прилагательному, особенно слабое склонение.	Наиболее частотная ЧР. Наибольшая коллизия с прилагательным и глаголом.	Использует 4-падежную парадигму. Морфологически близко к глаголу за счет ротацизма.	Наиболее частотная ЧР. Коллизия с глаголом, меньшей степени прилагательным.
Глаг.	Аблаут как механизм образования формы. Нефинитные формы используют адъективальную парадигму. Наиболее морфологически комплексная ЧР.	Наиболее точно и полно класси-фицируемая ЧР. Коллизия с существительным, в меньшей степени с прилагательным. Отрицательный глагол классифицирован	Аблаут как механизм образования формы. Нефинитные формы используют адъективальную парадигму. Наиболее морфологически комплексная ЧР.	Наиболее точно и полно класси-фицируемая ЧР. Коллизия с существительны м, незначительная с прилагательным.

		практически 100% точно.		
Прил.	4-падежная парадигма. Суперлатив совпадает с глагольным вторым лицом. Супплеция отдельных сравнительных форм. См. также «Сущ.» выше, «Нар». ниже.	Коллизия с существительным, в меньшей степени с наречием и глаголом. Коллизия с прочими ЧР практически отсутствует.	4-падежная парадигма. Компаратив совпадает со вторым лицом глагола. Некоторые суффиксы прилагательного совпадают с маркером медиопассива. Супплеция отдельных сравнительных форм.	Коллизия с существительным, в меньшей мере с глаголом, практически отсутствует коллизия с наречием.
Нар.	Не изменяется по форме. Отчасти использует суффиксы прилагательных.	Коллизия с существительным и прилагательным, также с предлогом (возможно, из-за препозиционных составных наречий).	Не изменяется по форме. Почти не использует суффиксы других ЧР.	Практически не коллидирует с другими ЧР.
Служ. и мест.	Значительная омонимия между различными частицами, предлогами, союзами, детерминативами.	Низкая точность классификации отдельных служебных слов. Местоимение *man* не классифи-цируется корректно.	Значимые особенности отсутствуют.	Точность и полнота классификации служебных слов выше, чем самостоятельных ЧР.
Прочие особ.	1 основной вид умлаута. Сильная орфографическая вариативность.	Более низкая точность и полнота классификации по сравнению с ДС.	3 вида умлаута. Дипломатическая запись нормализована орфографически.	Влияния умлаутов на точность и полноту классификации не выявлено.

ВЫВОДЫ ПО ВТОРОЙ ГЛАВЕ.

В результате анализа практического материала и проведения экспериментов можно сделать вывод о том, что к настоящему времени лингвисты особое внимание уделяют обработке текстов на естественных языках. Автоматическая частеречная разметка — задача, не разрешенная окончательно даже в отношении современного английского языка, и тем более, малоисследованная применительно к диахроническим вариантам германских языков, таким, как среднеанглийский или древнескандинавский.

Исходный размеченный корпусный материал (Parsed Linguistic Atlas of Early Middle English, Medieval Nordic Text Archive) потребовал существенной подготовки и переработки перед его векторизацией и снабжения векторов частеречными пометами в целях формирования обучающе-проверочной (за счет многопроходной перекрестной проверки) выборки. В частности, PLAEME использует комбинации символов стандартной латиницы и математических операторов в связи с тем, что предназначен для работы с устаревшей утилитой CorpusSearch, не поддерживающей кодировку UTF-8, в связи с чем у составителей корпуса возникла необходимость избавиться от специальных (в т.ч. рунических) символов. Данные сочетания были возвращены к исходному написанию, согласно технической документации корпуса. Удалены специальные пометы, характеризующие особый текст: надписанный, цитаты на латыни и т.д., при необходимости – вместе с самим текстом. Набор из 217 частеречных помет, зачастую встречающихся единично в проанализированном фрагменте корпуса, был сведен к 21 классу. Аналогичная процедура проведена в отношении древнескандинавского корпуса, где из более 100 оригинальных помет оставлено 11, при этом необходимости в восстановлении исходной орфографии не возникло, так как Menota – более современный корпус, и в нем используется кодировка UTF-8. Конечный текст обеих выборок методом скользящего среднего преобразован в матрицу, содержащую координаты вектор-слов в n-мерном пространстве, где n = число символов алфавита, умноженное на 3.

Экспериментальная проверка шести предложенных моделей машинного обучения показала, что четыре из них непригодны для использования в означенных целях, так как их точность и полнота неудовлетворительны. Два алгоритма, результаты которых достаточно точны (k ближайших соседей и метод случайного леса), на обоих языках демонстрируют схожие результаты (см. табл. 20). В качестве основных особенностей полученных выходных данных следует отметить:

- высокое качество (точность и полноту) классификации глагола;
- ошибочность классификации многих прилагательных, что обусловлено их морфографемической схожестью с глаголом, существительным и в меньшей степени – наречием;
- частые ошибки классификации отдельных служебных слов в среднеанглийском тексте одновременно с высокой точностью определения их классов в древнескандинавском корпусе, что обусловлено значительной омонимией в первом и ее отсутствием во втором;
- более высокие значения точности и полноты в отношении древнескандинавского текста.

Данные особенности указывают на одно из возможных направлений доработки реализованного в настоящем исследовании метода, а именно – устранение коллизии морфологически схожих именных частей речи (в первую очередь, прилагательного), снятие омонимии служебных слов. Подобная проблематика не представляется разрешимой в рамках исключительно посимвольного векторного кодирования и указывает на целесообразность дополнения выбранного метода репрезентации контекстуализированной векторизацией, например, с помощью n-грамм. Тем не менее, достигнутый результат считаем удовлетворительным, так как даже с учетом указанных недостатков достигнутая точность является высокой в отношении среднеанглийского языка, а в отношении древнескандинавского превышает то, что было достигнуто ранее другими исследователями.

Стоит отметить, что невзирая на кажущуюся схожесть результатов двух алгоритмов, точность и полнота случайного леса статистически достоверно выше аналогичных показателей метода k ближайших соседей, что подтверждается проверкой по критерию Стьюдента, практически по всем классам; данное обстоятельство означает, что объединение этих двух алгоритмов в ансамбль методом стекинга нецелесообразно. Ансамблирование kNN алгоритмом AdaBooster не является возможным ввиду специфики их реализации в использовавшемся инструментарии (библиотека sklearn); бустинг случайного леса, будучи нетривиальным и редко применяемым подходом, не обесечил статистически значимого повышения точности; кроме того, его вычислительная емкость позволяет рекомендовать подобное решение только в том случае, когда размер обучающе-тестирующей выборки относительно невелик (в данном случае – на выдержке из Menota).

Поскольку сам по себе случайный лес является ансамблирующим алгоритмом, считаем, что исходное положение данного исследования подтверждено, и оно позволило глубже проанализировать диахронический корпус сравниваемых языков.

ЗАКЛЮЧЕНИЕ

Среднеанглийский и древнескандинавский языки в тексте диахронического корпуса реализуют различные морфографемические особенности, за счет которых осуществляется различение частей речи. Древнескандинавский язык морфологически и фонетически сложнее: в нем присутствуют, помимо i-умлаута, a- и u-мутации, выполняюшие в том числе формообразовательную функцию, а также суффиксальный маркер определенности. Помимо умлаута, возникает так называемый аблаут – перебой корневой гласной как показатель прошедшего времени сильного глагола, а также некоторых личных форм претеритно-презентных глаголов в обоих языках. Таким образом, германские языки образуют словоформы не только префиксацией и суффиксацией, но и внутрикорневой флексией. Значимой проблемой представляется также частичное совпадение морфологических парадигм существительного и прилагательного, схожесть суперлативной формы прилагательного и второго лица глагола в среднеанглийском языке в связи с отсутствием ротацизма первой, а также компаративной формы и личных форм настоящего времени в древнескандинавском в связи с ротацизмом последних. Отчасти под влиянием французского языка среднеанглийский демонстрирует значительную вариативность орфографической реализации. Все перечисленные особенности затрудняют обработку текстов на подобном языке средствами ЭВМ.

В рамках компьютерной и корпусной лингвистики недоисследованным остается вопрос приложения методов машинного обучения к диахроническому языковому материалу, в частности, в вопросах автоматизации частеречного аннотирования – трудоемкого процесса, являющегося неотъемлемым шагом в подготовке исторического корпуса. Этим была обусловлена **цель** настоящего исследования: реализовать применение методов МО к диахроническому тексту и проверить их эффективность, выражаемую точностью и полнотой, с точки зрения классификации лексем по частям речи.

По итогам выполнения диссертационного исследования мы пришли к выводам, что орфографические особенности среднеанглийского и древнескандинавского языков, их морфологическая и фонетическая специфика, выявляемые сопоставительно-сравнительным анализом, оказывают влияние на качественные характеристики автоматического частеречного аннотирования корпуса текстов на древних языках. Обеспечение высокой точности аннотирования требует применения специального метода кодирования входных данных, в качестве такового в рамках настоящей работы был выбран метод скользящего среднего, предложенный П. Такалой. Метод позволил сформировать наборы вектор-слов относительно малой размерности (105 измерений для среднеанглийского и 141 для древнескандинавского языков), обеспечивающих морфографемическое посимвольное кодирование каждого слова с определением координаты в каждом измерении по весовым коэффициентам того или иного символа в слове, которые, в свою очередь, задавались его интралексемным положением.

Полученные данные были использованы в серии экспериментов, нацеленных на проверку точности и полноты шести различных алгоритмов машинного обучения: *k* ближайших соседей (kNN), модель случайного леса (RFM), метод опорных векторов, наивный баейсовский алгоритм, логистическая регрессия и многослойный перцептрон. Первые два алгоритма в отношении обоих языков показали наибольшую точность, при этом между kNN и RFM отмечена малая, но статистически значимая разница в пользу последнего по оцениваемым параметрам классификации всех классов. Остальные алгоритмы продемонстрировали низкую точность и полноту, в связи с чем были исключены из потенциальных участников метода-ансамбля. Стоит отметить, что независимо от алгоритма более высоких искомых значений удалось добиться на материале древнескандинавского языка, что мы связываем, с одной стороны, с его более устойчивой орфографией и более выраженной морфографемической дифференциацией частей речи, с другой – меньшей степенью омонимии служебных частей речи.

Всего два алгоритма показали удовлетворительный результат, и один из них оказался более точным и полным по сравнению с другим практически без исключений, от составления ансамбля методом стекинга, обеспечивающим предпочтительное (взвешенное) голосование алгоритмов-участников за счет подключения третьей, собственно взвешивающей модели, поэтому нами было решено от этого отказаться: стекинг является ресурсоемким решением, которое в данном случае не оказало бы положительного влияния на конечный результат. Другой тестируемый ансамбль – AdaBooster (ABC) – не применим к методу ближайших соседей, т.к. в реализации этого алгоритма в библиотеке sklearn языка Python отсутствует необходимый параметр sample_weight (весовой коэффициент выборки), без которого невозможен «бустинг» базового классификатора. Хотя использование RFM в качестве классификатора-основы в составе ABC является нетривиальным решением, т.к. оба алгоритма, по сути, являются ансамблями деревьев принятия решений с той разницей, что ABC реализует их последовательно, а RFM – параллельно, нами было принято решение осуществить «усиление» случайного леса как базового классификатора. Итоговый результат был достоверно лучше вывода RFM без усиления, однако более применимо к среднеанглийскому языку, поэтому, на наш взгляд, в рамках данного эксперимента подобный подход нельзя назвать оправданным ввиду значительной вычислительной ресурсоемкости ABC.

Таким образом, рекомендуемым решением классификации на основе обучения на векторизованной методом скользящего среднего выборки является применение случайного леса в отношении текста на среднеанглийском языке, случайного леса в сочетании с AdaBooster в отношении корпуса древнескандинавского языка. Необходимо отметить, что подобная рекомендация обусловлена исключительно размером использовавшихся в экспериментах выборок, а не спецификой самих языков. Невзирая на это, следует считать эмпирически доказанной выдвинутую в рамках исследования **гипотезу**: морфографемическое кодирование текста в рамках машинного обучения вкупе с

ансамблированием обеспечивает высокоэффективную автоматизацию процесса частеречной разметки при обучении с учителем на малой выборке.

Результаты исследования релевантны в первую очередь в области исторической лингвистики, так как закладывают основы создания инструмента упрощенной, быстрой и точной обработки диахронического текста; корпусной и компьютерной лингвистики, так как, задействуя относительно простое с математической точки зрения решение, позволяют добиться высокой по сравнению с имеющимися выводами других исследователей точности и полноты классификации по частеречному признаку.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Арапов М.А., Херц М.М. Математические методы в исторической лингвистике. Москва: Наука, 1973. 322 с.
2. Артюхин В.В. Применение методов машинного обучения при работе с литературными источниками // Образовательные технологии и общество. 2015. Т. 18. № 2. С. 601–617.
3. Асташкин А.Г., Чувилин К.В. Генерация описания синтаксиса с помощью методов машинного обучения // Труды Международной научной конференции CPT1617. Москва: ИФТИ, 2017. С. 261–266.
4. Большакова Е.И. и др. Автоматическая обработка текстов на естественном языке и анализ данных. Москва: НИУ ВШЭ, 2017. 269 с.
5. Гапонов Д.А. Использованиме методов машинного обучения для тематической классификации текстовых данных // Труды молодых ученых Алтайского государственного университета. 2017. № 14. С. 243–246.
6. Гущин А.Е. Методы ансамблирования обучающихся алгоритмов: выпускная квалификационная работа бакалавра. Москва, 2015.
7. Доля П.Г. Введение в научный Python. Харьков: Харьковский национальный университет, 2016. 265 с.
8. Заковоротная Е.М., Северина Е.В. Использование параллельных корпусов в системах автоматизированного перевода // Научные исследования и разработки 2018 года. Новосибирск: Центр развития научного сотрудничества, 2018. С. 88–96.
9. Захаров В.П. Корпусная лингвистика // Прикладная и компьютерная лингвистика / под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. Москва: Ленанд, 2016. С. 140–157.
10. Иванова И.П., Чахоян Л.П., Беляева Т.М. История английского языка. Москва: Азбука-классика, 2006. 560 с.

11. Ильиш Б.А. История английского языка. Ленинград: Просвещение, 1973. 351 с.
12. Каримов Р.Д., Акинин А.А. Расширенный редактор корпуса с графическим пользовательским интерфейсом // Вестник ПНИПУ. Проблемы языкознания и педагогики. 2017. № 2. С. 67–76.
13. Климов Д.В. Редукция признакового пространства матрицы объектов в задаче классификации текстов // Современные условия взаимодействия науки и техники. Омск: ОМЕГА САЙНС, 2017. С. 79–82.
14. Краснопеева Е.С. Лексические особенности русскоязычного переводного дискурса (корпусное сравнительно-сопоставительное исследование на материале современной художественной прозы): дис. … канд. филол. наук. Челябинск, 2015. 202 с.
15. Молдагулова А.М., Бектемысова Г.У. Разработка системы с использованием методов машинного обучения для компьютерной лингвистики и обработки естественных языков // Актуальные вопросы современной науки. Уфа: Дендра, 2017. С. 23–30.
16. Молошников И.А. и др. Двухуровневая модель нейронной сети глубокого обучения для задачи морфологического разбора предложений русского языка // Вестник НИЯУ МИФИ. 2017. Т. 6. № 6. С. 555–562.
17. Найденова К.А., Невзорова О.А. Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований // Ученые записки Казанского университета. Серия Физико-математические науки. 2008. Т. 150. № 4. С. 5–22.
18. Николаева Н.А. Тематизация презенса сильного глагола в кельтских и германских языках: дис. … канд. филол. наук. М., 2003. 200 с.
19. Новиков Г.Г., Ядыкин И.М. Модуль нормализации корпуса текстов для метода Doc2Vec. 2016.

20. Полицын С.А., Полицына Е.В. Применение комплекса инструментов управления корпусами текстов при решении задач компьютерной лингвистики // Вестник Воронежского государственного университета. Серия Системный анализ и информационные технологии. 2019. № 2. С. 134–142.
21. Попков М.И. Автоматическая классификация текстов для базы знаний предприятия // Int. J. Open Inf. Technol. 2014. Т. 2. № 7. С. 11–18.
22. Расторгуева Т.А. История английского языка. Москва: АСТ, 2003. 348 с.
23. Родионова Е.С. Лингвистические методы атрибуции и датировки литературных произведений: к проблеме «Корнель - Мольер». 2008.
24. Стеблин-Каменский М.И. История скандинавских языков. Москва: Академия наук СССР: Институт языкознания, 1953. 337 с.
25. Тарасов Д.В., Романов Н.А. Процедура машинного обучения в задаче морфологической разметки текста и определения частей речи в флективных языках // Известия высших учебных заведений. Поволжский регион. Технические науки. 2017. № 1 (41). С. 56–72.
26. Что такое Custom Translator? - Azure Cognitive Services | Microsoft Docs [Электронный ресурс]. URL: https://docs.microsoft.com/ru-ru/azure/cognitive-services/translator/custom-translator/overview
27. Шелманов А.О., Каменская М.А. Обучение анализатора для определения ролевых структур высказываний в текстах на русском языке на автоматически размеченном корпусе // Труды Института системного анализа Российской Академии наук. 2017. Т. 67. № 2. С. 104–120.
28. Шенбин И.И. Обучение распределенных представлений слов на основе символов: магистерская диссертация. СПб, 2017.
29. Шестаков К.М. Теория принятия решений. Москва: МГУ, 2012. 182 с.
30. Aha D.W., Kibler D., Albert M.K. Instance-Based Learning Algorithms // Mach. Learn. 1991. Vol. 6. No 1. Pp. 37–66.

31. Backus J.W. The syntax and semantics of the proposed international algebraic language 125 c The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM conference // Proceedings of the International Conference on Information Processing. UNESCO, 1959. Pp. 125–132.
32. Backus J.W. et al. Revised report on the algorithmic language Algol 60 // Numer. Math. 1962.
33. Bandle O. The Nordic languages: an international handbook of the history of the North Germanic languages. Berlin: Walter de Gruyter, 2002. 1050 p.
34. Barteld F., Schröder I., Zinsmeister H. Unsupervised regularization of historical texts for POS tagging // Proceedings of the 4th Workshop on Corpus-based Research in the Humanities. Warsaw: 2015. Pp. 3–12.
35. Berger A.L., Pietra V.J. Della, Pietra S.A. Della. A Maximum Entropy Approach to Natural Language Processing // Comput. Linguist. 1996. Vol. 22. No 1. Pp. 39–71.
36. Bledsoe W.W., Browning I. Pattern recognition and reading by machine // Papers presented at the December 1-3, 1959, eastern joint IRE-AIEE-ACM computer conference on - IRE-AIEE-ACM '59 (Eastern). New York, New York, USA: ACM Press, 1959. Pp. 225–232.
37. Blei D.M., Ng A.Y., Edu J.B. Latent Dirichlet Allocation Michael I. Jordan // J. Mach. Learn. Res. 2003. Vol. 3. Pp. 993–1022.
38. Boser B.E., Guyon I.M., Vapnik V.N. A Training Algorithm for Optimal Margin Classiiers // COLT. 1992. Pp. 144–152.
39. Breiman L. Random Forests. Berkeley: 2001. 33 p.
40. Bresnan J., Kaplan R.M. Grammars as mental representations of language. Boston: MIT Press, 1982. 500 p.
41. Brown P.F. et al. A statistical approach to machine translation // Comput. Linguist. 1990. Vol. 16. No 2. Pp. 79–85.
42. Brunner K., Johnson G. An outline of Middle English grammar. Oxford: Basil Blackwell, 1970. 128 p.

43. Buys J., Botha J.A. Cross-Lingual Morphological Tagging for Low-Resource Languages // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: 2016. Pp. 1954–1964.
44. Campbell L. Historical Linguistics. Edinburgh: Edinburgh University Press, 2013. Iss. 3. 538 p.
45. Cao K., Rei M. A Joint Model for Word Embedding and Word Morphology // Comput. Sci. Comput. Linguist. 2016.
46. Carbonell J.G., Cullingford R.E., Gershman A. V. Steps Toward Knowledge-Based Machine Translation // IEEE Trans. Pattern Anal. Mach. Intell. 1981.
47. Carlson L. et al. Discourse Tagging Reference Manual. 2001.
48. Celano G.G.A., Crane G., Majidi S. Part of speech tagging for ancient Greek // Open Linguist. 2016.
49. Cellier P. et al. Using Machine Learning Methods and Linguistic Features in Single-Document Extractive Summarization. 2016.
50. Chomsky N. A Review of B. F. Skinner's Verbal Behavior // Language (Baltim). 1959. No 1. Pp. 26–58.
51. Chomsky N. Three models for the description of language // IRE Trans. Inf. Theory. 1956.
52. Church K.W. On memory limitations in natural language processing // 1980.
53. Clackson J. Indo-European Linguistics: an Introduction. Cambridge: Cambridge University Press, 2007. 246 p.
54. Colmerauer A. Metamorphosis Grammars // Natural Language Communication with Computers. 2006. Pp. 133–188.
55. Corpuscle Documentation [Electronic resource]. URL: http://clarino.uib.no/menota/page?page-id=korpuskel-documentation
56. CorpusSearch | Home [Electronic resource]. URL: http://corpussearch.sourceforge.net/

57. Cristianini N., Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, 2000. 189 p.
58. Daelemans W., Durieux G., Gillis S. The Acquisition of Stress: A Data-Oriented Approach // Comput. Linguist. 1994. Vol. 20. No 3. Pp. 421–451.
59. Dalton-Puffer C. The French influence on Middle English morphology : a corpus-based study of derivation. Mouton de Gruyter, 1996. 284 p.
60. Dang H.T. Overview of DUC 2006. 2006.
61. Davis K.H., Biddulph R., Balashek S. Automatic Recognition of Spoken Digits // J. Acoust. Soc. Am. 1952. Vol. 24. No 6. Pp. 637–642.
62. Derczynski L. et al. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. 2013. 7–13 p.
63. Diederich J. et al. Authorship attribution with support vector machines // Appl. Intell. 2003.
64. Dilts P. Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed- Effects Regression // 2013.
65. Du K.-L., Swamy M.N.S. Multilayer Perceptrons: Architecture and Error Backpropagation // Neural Networks and Statistical Learning. London: Springer London, 2014. Pp. 83–126.
66. Faarlund J.T. The syntax of Old Nordic // The Nordic Languages: An International Handbook of the History of the North Germanic Languages. 2008. Pp. 940–950.
67. Fillmore C.J. The Case for Case // Universals in Linguistic Theory / Eds. E.W. Bach, R.T. Harms. Holt, Rinehart & Winston, 1968. Pp. 1–88.
68. Genkin A., Lewis D.D., Madigan D. Large-Scale Bayesian Logistic Regression for Text Categorization // Technometrics. 2007. Vol. 49. Pp. 291–304.
69. Glickman L. V., Rowntree D. Statistics without Tears, a Primer for Non-Mathematicians // Math. Gaz. 1982.

70. Gries S.T. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement // Corpus Linguist. Linguist. Theory. 2015. Vol. 78. Pp. 1–31.
71. Grosz B. The Representation and Use of Focus in a System for Understanding Dialogs. 1977. Pp. 67–76.
72. Guarasci R. Developing an annotator for Latin texts using Wikipedia // J. Data Min. Digit. Humanit. 2017.
73. Guo G. et al. An kNN Model-Based Approach and Its Application in Text Categorization. Springer, Berlin, Heidelberg, 2004. Pp. 559–570.
74. Hackeling G. Mastering Machine Learning with scikit-learn. Birmingham: Packt Publishing, 2014. 220 p.
75. Haeberli E., Ingham R. The position of negation and adverbs in Early Middle English // Lingua. 2007. Vol. 117. No 1. Pp. 1–25.
76. Haegstad M., Torp A. Gamalnorsk ordbok : med Nynorsk tyding. Norske Samlaget: Norske Samlaget, 1909. 652 p.
77. Hagland J.R. A Note on Old Norwegian Vowel Harmony // Nord. J. Linguist. 1978. Vol. 1. No 2. Pp. 141–147.
78. Hartshorn S. Bayes Theorem Examples: A Visual Guide For Beginners. Amazon Digital Services LLC, 2016. 82 p.
79. Hasan K.S., Ng V. Weakly supervised part-of-speech tagging for morphologically-rich, resource-scarce languages. 2009.
80. Haugen O.E. Grunnbok i norrønt språk. Gyldendal Akademisk, 1995. 320 p.
81. Haugen O.E. Handbok i norrøn filologi. Bergen: John Grieg AS, 2013. Iss. 2. 796 p.
82. Haugen O.E. Norrøne tekster i utval. Oslo: Gyldendal Akademisk, 1994. 312 p.
83. Hemming C. Using Neural Networks in Linguistic Resources. Skovde: 2003. 11 p.

84. Hobbs J.R. Resolving pronoun references // Lingua. 1978. Vol. 44. No 4. Pp. 311–338.
85. Hogg R. An Introduction to Old English. Edinburgh: Edinburgh University Press, 2002. 138 p.
86. Horobin S., Smith J. An Introduction to Middle English. Edinburgh: Edinburgh University Press, 2002. 182 p.
87. Iacobini C., Rosa A. De, Schirato G. Part-of-Speech tagging strategy for MIDIA: a diachronic corpus of the Italian language // Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014. Pisa: Pisa University Press, 2014. Pp. 213–218.
88. Jahr E.H., Lorentz O. Historisk språkvitenskap = Historical linguistics. Novus, 1993. 431 p.
89. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition // Speech Lang. Process. An Introd. to Nat. Lang. Process. Comput. Linguist. Speech Recognit. 2009.
90. Kaplan R.M., Kay M. Phonological rules and finite--state transducers // 1981.
91. Karimov R. Combined Machine--Learning Approach to PoS-Tagging of Middle English Corpora // Crossroads. A J. English Stud. 2018. Vol. 21. No 2. Pp. 42–52.
92. Karimov R., Akinin A., Yakymets D. Combined machine-learning approach to PoS-tagging of Middle English and Old Norse Texts // CEUR Workshop Proceedings. 2018.
93. Karttunen L. Applications of Finite-State Transducers in Natural Language Processing. 2001. Pp. 34–46.
94. Kay M. Functional grammars // Proceedings of the Berkeley Linguistics Society. Berkeley: 1979. Pp. 142–158.
95. Kilgarriff A., Palmer M. Introduction to the special issue on SENSEVAL // Comput. Hum. 2000. Vol. 34. Pp. 1–13.

96. Kleene S.C. Representation of Events in Nerve Nets and Finite Automata // Automata Studies. (AM-34). 2016.
97. Koenig W., Dunn H.K., Lacy L.Y. The Sound Spectrograph // J. Acoust. Soc. Am. 1946. Vol. 18. No 1. Pp. 19–49.
98. Koleva M. et al. An automatic part-of-speech tagger for Middle Low German // Int. J. Corpus Linguist. 2017. Vol. 22. No 1. Pp. 107–140.
99. König E., Auwera J. van der. The Germanic languages. Routledge, 2002. 631 p.
100. Kroch A., Taylor A. Penn Parsed Corpora of Historical English [Electronic resource]. URL: https://www.ling.upenn.edu/hist-corpora/
101. Kučera H., Nelson F.W. Computational analysis of present-day American English. Providence: Brown University Press, 1967. 200 p.
102. Laing M. A Linguistic Atlas of Early Middle English, 1150–1325 // 2013.
103. Larsson F. Centre for Language and Literature Two Millennia of Lexical and Typological Change in Western Europe-a quantitative geographical approach. Lund: 2013. 68 p.
104. Lehnert W.G. A conceptual theory of question answering // Proceedings of the 5th international joint conference on Artificial intelligence - Volume 1. 1977. Pp. 158–164.
105. Leopold E., Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? // Mach. Learn. 2002. Vol. 46. No 1/3. Pp. 423–444.
106. Leshem G. Improvement of Adaboost Algorithm by using Random Forests as Weak Learner and using this algorithm as statistics machine learning for traffic flow prediction. Jerusalem: 2005. 30 p.
107. Loftsson H., Helgadóttir S., Rögnvaldsson E. Using a Morphological Database to Increase the Accuracy in {POS} Tagging // Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. Hissar, Bulgaria: Association for Computational Linguistics, 2011. Pp. 49–55.

108. Longacre R.E., Harris Z.S. String Analysis of Sentence Structure // Language (Baltim). 2006.
109. Magnus R. Norsk språk 1350–1500. Gammalnorsk eller mellomnorsk? // Historisk språkvitenskap/Historical Linguistics / Eds. E.H. Jahr, O. Lorentz. 1993. Pp. 395–404.
110. Manning C.D. et al. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2012. 54 p.
111. Manning C.D., Schütze H. Foundations of Natural Language Processing // Reading. 2000.
112. Marco C.S. An open source part-of-speech tagger for Norwegian : Building on existing language resources // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). Reykjavik: European Languages Resources Association (ELRA), 2012. Pp. 4111–4117.
113. Mayhew A.L. (Anthony L., Skeat W.W. (Walter W., Everson M. A concise dictionary of Middle English : from 1150 to 1580. Evertype, 2011. 272 p.
114. McCulloch W.S., Pitts W. A logical calculus of the ideas immanent in nervous activity // Bull. Math. Biophys. 1943.
115. McTurk R. A Companion to Old Norse-Icelandic Literature. Blackwell Publishing Ltd, 2005. 534 p.
116. Medieval Nordic Text Archive (Menota) [Electronic resource]. URL: https://menota.org/forside.
117. Mikolov T. et al. Efficient Estimation of Word Representations in Vector Space // Comput. Sci. Comput. Lang. 2013. Pp. 1–12.
118. Millward C.M., Hayes M. A Biography of the English Language. Boston: Wadsworth, 2012. 478 p.
119. Miltsakaki E. et al. Annotating Discourse Connectives and Their Arguments // 2004. Pp. 9–16.

120. Moon T., Baldridge J. Part-of-speech tagging for middle English through alignment and projection of parallel diachronic texts // Proc. 2007 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. 2007.
121. Mosteller F., Wallace D.L. Inference in an Authorship Problem // J. Am. Stat. Assoc. 1963.
122. Mustanoja T.F. (Tauno F., Gelderen E. van. A Middle English syntax. Amsterdam: John Benjamin's Publishing Company, 2016. 702 p.
123. Narayanan I. et al. SSD Failures in Datacenters: What? When? and Why? // SYSTOR '16. Haifa, Israel: ACM, 2016.
124. Narayanan S., Narayanan S., Jurafsky D. Bayesian Models of Human Sentence Processing // PROCEDINGS 20 TH Annu. Conf. Cogn. Sci. Soc. 1998. Pp. 752–757.
125. Ngo K.T. Stacking Ensemble for auto ml. 2018.
126. Nogueira C., Santos D., Zadrozny B. Learning Character-level Representations for Part-of-Speech Tagging // Proceedings of the 31st International Conference on Machine Learning. 2014.
127. Norman D.A. et al. Explorations in Cognition. LNR Research Group, 1975. 430 p.
128. Oparin I., Lamel L., Gauvain J.-L. Large-Scale Language Modeling with Random Forests for Mandarin Chinese Speech-to-Text // Advances in Natural Language Processing. 2010. Pp. 269–280.
129. Padró L. Analizadores Multilingües en FreeLing // Linguamatica. 2011. Vol. 3. No 2. Pp. 13–20.
130. Palmer M., Gildea D., Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles // Comput. Linguist. 2005. Vol. 31. No 1. Pp. 71–106.
131. Pearl J., Judea. Probabilistic reasoning in intelligent systems : networks of plausible inference. Morgan Kaufmann Publishers, 1988. 552 p.

132. Pereira F.C.N., Warren D.H.D. Definite clause grammars for language analysis-A survey of the formalism and a comparison with augmented transition networks // Artif. Intell. 1980.
133. Perrault C.R. A Plan-Based Analysis of Indirect Speech Act // Am. J. Comput. Linguist. 1980. Vol. 6. No 3–4. Pp. 167–182.
134. Pranckevicius T., Marcinkevicius V. Application of Logistic Regression with part-of-the-speech tagging for multi-class text classification // 2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE). IEEE, 2016. Pp. 1–5.
135. Pustejovsky J. et al. The TimeBank corpus // Proc. Corpus Linguist. 2003. Vol. 1. Pp. 647–656.
136. Quillian M.R. Semantic Memory // Semantic Information Processing / Ed. M. Minsky. Boston: MIT Press, 1968. Pp. 227–270.
137. Rabiner L., Juang B.-H. Fundamentals of Speech Recognition. Englewood Cliffs: Prentice Hall, 1993. 497 p.
138. Ralph W.V., Elliott M.A. Runes: an Introduction. Manchester University Press, 1963. 117 p.
139. Rask E. A Grammer of the Icelandic or Old Norse Tongue. London: William Pickering, 1843. 272 p.
140. Ratnaparkhi A. A Maximum Entropy Model for Part-Of-Speech Tagging // EMNLP. 1996. Pp. 133–142.
141. Sadov M.A., Kutuzov A.B. Use of morphology in distributional word embedding models: Russian language case. 2018.
142. Sang E.F.T.K., Dejean H. Introduction to the CoNLL-2001 Shared Task: Clause Identification // Proceedings of CoNLL-2001. Toulouse: 2001. Pp. 53–57.
143. Schank R.C., Riesbeck C.K., Riesbeck C.K. Inside Computer Understanding. Psychology Press, 2017. 400 p.

144. Schulte M. Die lateinisch-altrunische Kontakthypothese im Lichte der sprachhistorischen Evidenz // Beiträge zur Geschichte der Dtsch. Spr. und Lit. 2005. Vol. 127. No 2. Pp. 161–182.
145. Shannon C.E. A Mathematical Theory of Communication // Bell Syst. Tech. J. 1948.
146. Silva A.P., Silva A., Rodrigues I. An Approach to the POS Tagging Problem Using Genetic Algorithms. Springer, Cham, 2015. Pp. 3–17.
147. Simmons R.F. Semantic Networks: Their Computation and Use for Understanding English Sentences // Computer Models of Thought and Language / Eds. R.C. Schank, K.M. Colby. W.H.Freeman & Co., 1973. Pp. 61–113.
148. Smit P. et al. Morfessor 2.0: Toolkit for statistical morphological segmentation // Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014. Pp. 21–24.
149. Sokolova M., Japkowicz N., Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation // AI 2006: Advances in Artificial Intelligence, 19th Australian Joint Conference on Artificial Intelligence / Eds. A. Sattar, Byeong Ho Kang. Hobart: tralian Computer Society, 2006. Pp. 1015–1021.
150. Sonawane J.S., Patil D.R. Prediction of heart disease using multilayer perceptron neural network // International Conference on Information Communication and Embedded Systems (ICICES2014). IEEE, 2014. Pp. 1–6.
151. Spurkland T. Innføring i norrønt språk. Oslo: Unversitetsforlaget, 1989. 173 p.
152. Stein A. Syntactic Annotation of Old French Text Corpora // Corpus. 2008. No 7. Pp. 157–171.

153. Stubbs W., Davis H.W.C. Select charters and other illustrations of English constitutional history, from the earliest times to the reign of Edward the First. Oxford: Clarendon Press, 1921. 556 p.
154. Taddy M. Multinomial Inverse Regression for Text Analysis // J. Am. Stat. Assoc. 2013. Vol. 108. Pp. 755–770.
155. Takala P. Word Embeddings for Morphologically Rich Languages // 2016. No April. Pp. 27–29.
156. Taylor A., Marcus M., Santorini B. The Penn Treebank: An Overview. 2003.
157. Tharwat A. AdaBoost classifier: an overview. Frankfurt: 2018. 72 p.
158. The importance of Data Representation - Data Mining and Reporting [Electronic resource]. URL: https://sites.google.com/site/dataminingandreporting/news/data-mining/the-importance-of-data-representation.
159. Tripathi N., Oakes M., Wermter S. A Fast Subspace Text Categorization Method Using Parallel Classifiers // Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II / Eds. A.A. Gelbukh. Berlin: Springer-Verlag, 2012. Pp. 132–143.
160. Trosterud T. The changes in Scandinavian morphology from 1100 to 1500 // Ark. för Nord. Filol. 2001. No 116. Pp. 159–191.
161. Truswell R. et al. A Parsed Linguistic Atlas of Early Middle English.
162. Vapnik V.N. The nature of statistical learning theory. Springer, 2000. 314 p.
163. VOORHEES E.M., M. E. The TREC question answering track // Nat. Lang. Eng. 2001. Vol. 7. No 4. Pp. 361–378.
164. Wilks Y. A preferential, pattern-seeking, Semantics for natural language inference // Artif. Intell. 1975. Vol. 6. No 1. Pp. 53–74.
165. Winograd T. Understanding natural language // Cogn. Psychol. 1972. No 3(1).

166. Woods W.A. Semantics for a question-answering system. New York: Garland Pub, 1979. 346 p.
167. Wyner A.J. et al. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers // J. Mach. Learn. Res. 2017. No 18. Pp. 1–33.
168. Yang Y., Eisenstein J. Part-of-Speech Tagging for Historical English // NAACL-HLT 2016. San Diego: Association for Computational Linguistics, 2016. Pp. 1318–1328.
169. Yergeau F. RFC 3629 - UTF-8, a transformation format of ISO 10646. 2003. 15 p.
170. Young T. et al. Recent Trends in Deep Learning Based Natural Language Processing [Review Article] // IEEE Comput. Intell. Mag. 2018. Vol. 13. No 3. Pp. 55–75.

ПРИЛОЖЕНИЕ А. Пример текста из корпуса PLAEME.

Исходные файлы корпуса с разметкой, адаптированной для использования в утилите CorpusSearch, доступны по ссылке: https://datashare.is.ed.ac.uk/handle/10283/3032

Worldes blis ne last no þrowe . hit wit
ant wend a-wey a-non . þe lengur þat
hich hit i-knowe . þe lasse hic finde pris
þer-on . for al hit is imeynd wyd kare .
mid sorewe ant wid uuel fare . ant at
þe laste pouere ant bare hit let mon wen hit ginnet a-gon. al þe blisse
þis here ant þere bi-louketh at hende wop ant Mon .
Al shal gon þat her mon howet . al hit shal wenden to
nout . þe mon þat her no god ne sowet . wen oþer repen he
worth bikert . þenc mon for-þi wil þu hauest mykte . þat þu
þine gultus here arikte ant wrche god bi day an nikte . ar
-þen þu be of lisse ilakt . þu nost wanne crist ure drikte .
þe asket þat he hauet bitakt . Al þe blisse of þisse liue . þu
shalt mon henden in wep . of huse ant home ant child ant
wyue . seli mon tak þer-of kep . for þu shalt al bileuen here .
þe eykte were-of louerd þu were . wen þu list mon
up-on bere . ant slepest a swyþe druye slep . ne shaltu haben
wit þe no fere . butte þine werkus on an hep . Mon wi
seestu loue ant herte . on worldes blisse þat nout ne last .
wy þolestu þat te so ofte smerte . for loue þat is so unstedefast .
þu likest huni of þorn iwis . þat seest þi loue on worldes blis .
for ful of bitternis hit is . sore þu mikt ben of-gast . þat despen
des here heikte amis wer-þurk ben in-to helle itakt . Þenc
mon war-of crist þe wroukte . ant do wey prude ant fulthe

mod . þenc wou dere he þe bokte . on rode mit his swete blod .
him-self he gaf for þe in pris . to buge þe blis yf þu be wis .
bi-þenc þe mon ant up-aris . of slovþe an gin to worche god .
wil time to worchen is . for elles þu art witles ant wod .
Al day þu mikt undurstonde
ant ti mirour bi-for þe sen .
wou þis world wat is to don an to wonden ant wat to hol
den ant to flen . for al day þu sigst wid þin egven wou þis
world went . ant wou men deiegt . þat wite wel þat þu shalt
þu dreigen det al-so an-oþer det . ne helput nout þer non to
ligen . ne may no-mon bu det ageyn . Ne wort ne god
þer unforgulde ne non uuel ne worth un-bokt wanne
þu list mon undur molde . þu shalt hauen astu hauest
wrokt . bi-þenc þe wel for-þi hic rede . ant clanse þe of þine
misdede . þat he þe helpe at þine nede . þat so dure hus haued
iboukt . ant to heuene blisse lede . þat euere lest ant failet nout .

ПРИЛОЖЕНИЕ Б. Текст Konungs skuggsjá, использовавшийся как образец древнескандинавского языка.

Полный текст в дипломатической записи с отображением разметки доступен по адресу: http://clarino.uib.no/menota/document-element?session-id=247665532348828&cpos=105480&corpus=menota

Sunr

18 Goðan dag
19 hærra min*n*.
20 Ec em sꝛ a ko·
21 min*n* til yðars
22 fun*n*dar sæm
23 byriar lyðnum
24 syni oc litillato*m*
25 at fin*n*a astsam·
26 legan foður oc gofgan. Oc ꝛ il ec þæs
27 yðr biðia at þer hafer þolenmœðe at
28 lyða þeim lutum er mec fyser at
29 spyria. en litillæti at sꝛ ara mæð goð ·
30 fysi spurnu*m* lutu*m*.

Faðeʀ

18 Mæð þꝛ i at þu ert æinka son min*n*
19 þa licar mer ꝛ æl at þu komer opt
20 til mins fun*n*dar. þꝛ i at skyllt ꝛ æri tal
21 occat um marga luti oc ꝛ il ec giarna ly·
22 ða þꝛ i er þu ꝛ ilt spurt hafa oc ꝛ æita
23 an*n*sꝛ or þei*m* lutu*m* er ec em mæð skynsæmð ˋspurðr.ˊ

Sunr

24 Ec heyri þ*at* alþyðu ·
25 ꝛ itni sæm ec ætla satt ꝛ æra
26 um ꝛ itzku yðra at ꝛ arla fai ꝛ itrari man*n*
27 i þæsso lan*n*de en þer erut til alrar spæki
28 þꝛ i at til yðarrar orlausnar stun*n*da aller
29 þæir er ꝛ an*n*da malu*m* æigu at skipta.
30 En*n*da heyri ec oc at sꝛ a ꝛ æri þa er þer ꝛ aruð

ПРИЛОЖЕНИЕ В. Список сокращений.

ABC — AdaBoost classifier, классификатор AdaBoost.

KNN — k nearest neighbors, k ближайших соседей.

LAEME — Linguistic Atlas of Early Middle English, Лингвистический атлас раннего среднеанглийского языка.

LR — logistic regression, логистическая регрессия.

MLP — multilayered perceptron, многослойный перцептрон.

NB — Naïve Bayes, наивный байесовский классификатор.

PLAEME — Parsed Linguistic Atlas of Early Middle English, Аннотированный Лингвистический атлас раннего среднеанглийского языка.

RFM — random forest model, модель случайного леса.

SVM — support vector machine, метод опорных векторов.