Рауль Дамирович КАРИМОВ[1]

УДК 81-112

ХАРАКТЕРИСТИКА ДРЕВНЕСКАНДИНАВСКОГО ЯЗЫКА С ТОЧКИ ЗРЕНИЯ АВТОМАТИЗАЦИИ ЧАСТЕРЕЧНОЙ РАЗМЕТКИ

[1] аспирант кафедры романо-германских языков
и межкультурной коммуникации,
Челябинский государственный университет
raoul.karimov@hotmail.com; ORCID: 0000-0003-0313-0309

Аннотация

В настоящей статье рассматривается проблема частеречной разметки древнескандинавского языка средствами ЭВМ, в том числе машинного обучения, с позиции исторического языкознания. Анализируются диахронические особенности исследуемого языкового материала с точки зрения их влияния на качество осуществляемой автоматизации процесса внесения такой разметки. Описывается характер фонетических аспектов языка, обусловивших возникшие ошибки классификации.

В качестве материала исследования используется текст древненорвежского трактата Konungs skuggsjá, векторизованный методом скользящего среднего, затем примененный для обучения модели случайного леса, усиленной алгоритмом AdaBoost. Моделирование обеспечивает высокую выходную точность порядка 97%. Не будучи контекстуально уточненной, применяемая векторизация не обеспечивает полное различение морфологически схожих частей речи: глагола, существительного, прилагательного и наречия. На это указывают как определенные в качестве ключевых параметров классификации векторные измерения, каждое из которых соответствует определенному символу, так и выделенные алгоритмом Morfessor наиболее частотные морфы. Анализ этих морфов позволяет определить перечень морфограмматических единиц, вызывающих наибольшее число ошибок классификации.

Цитирование: Каримов Р. Д. Характеристика древнескандинавского языка с точки зрения автоматизации частеречной разметки / Р. Д. Каримов // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2019. Том 5. № 4 (20). С. 38-52.
DOI: 10.21684/2411-197X-2019-5-4-38-52

Рассматривая выделенные морфы в историческом аспекте, отмечаем, что их коллизия обусловлена наследованием аналогично схожих морфов из протогерманского языка в контексте процесса, известного как ротацизм, т. е. преобразования ПГ /z/ в древнескандинавский /r/. Однако тот же самый процесс позволяет избежать коллизии личных глагольных форм, подвергшихся ротацизму, и родительного падежа существительных, унаследовавшего протогерманское окончание *-s*.

Основной вывод заключается в том, что, ввиду неизбежности морфологической коллизии, посимвольной векторной репрезентации может оказаться недостаточно при обучении на малой выборке или при постановке задачи по различению не только частей речи, но и словоформ.

Ключевые слова

Древнескандинавский язык, корпусная лингвистика, фонетика, морфология, части речи, разметка, векторное представление.

DOI: 10.21684/2411-197X-2019-5-4-38-52

Введение

Сравнительно-историческое языкознание на сегодняшний день можно назвать одним из тех разделов лингвистической науки, где в значительной мере используется *корпус* — специально размеченный массив текстовых данных, структура и архитектура которого приспособлены для машинного анализа в рамках какой-либо специальной дисциплины. О востребованности исторического или диахронического корпуса говорят такие факты, как привлечение крупных лингвистических коллективов к их созданию (в частности, Хельсинкский корпус английского языка создавался командой из 27 лингвистов, программистов и студентов-филологов [10]), равно как и популярность такого рода ресурсов на корпусных платформах наподобие SketchEngine, где вебкорпус английского языка с диахроническими пометами является самым наполненным, см. рис. 1.

Применимость корпусных технологий в исторической лингвистике широка: это и глоттохронологическое исследование эволюции количественных характеристик словаря [1], и хронологический анализ какого-либо отдельного историко-грамматического процесса в сопоставительно-сравнительном изучении нескольких языков, например, становления тематического спряжения сильного глагола [2], и изучение гармонизации гласных [5], и культурологический анализ текста в целях описания общественных явлений в историческом контексте, таких как мужская гомосексуальность в древнескандинавских племенах [4].

Одним из основных видов разметки, применяемых в корпусе и критичных в историко-лингвистическом анализе, является частеречное аннотирование (англ. PoS-tagging) [13]. При этом, несмотря на разнообразие и большое количество имеющихся исторических корпусов, у исследователя может возникнуть потребность в создании собственного, в частности, в целях внесения в него текста,

который отсутствует в существующих корпусах. При этом ручная разметка может оказаться слишком трудоемкой и ресурсозатратной, в связи с чем при разработке корпуса часто обращаются к средствам автоматизации, например, алгоритмам TreeTagger, TNT или конечным автоматам [9]; однако для таких алгоритмов отсутствуют модели, обученные на материале исторических языков, в частности древнескандинавского. В связи с этим представляется актуальным проанализировать древнескандинавский язык как объект и материал машинного обучения анализаторов-классификаторов, применимых для решения подобных задач.

Материал и методика исследования

Настоящее исследование выполнено на материале древнескандинавского языка, само определение которого несколько спорно. О. Бэндл [3] формулирует общий консенсус: древнескандинавский диалектный континуум охватывал территорию от юга и запада Норвегии (а начиная с конца IX в. — еще и Исландии), Фарер и Шетландских островов до Ютландии на юге и даже юго-запада России на востоке. Отдельные исследователи заявляют, что только западные диалекты (норвежские и исландские) можно определять как Old Norse, другие называют западные диалекты общим термином «древненорвежский язык», хотя здесь возникает терминологическая коллизия, так как слово *gammalnorsk* чаще используют для обозначения уже самостоятельного, четко изолируемого от исландского, языка Норвегии периода 1350-1500 гг. (однако тот же язык именуют «средненорвежским», или *mellomnorsk* [8]). Однозначно можно выделить три ключевых диалектных континуума: западный и восточный древнескандинавские языки, а также гутнийский — диалект, изоглоссируемый по границе Гётланда, Швеция. В рамках настоящей работы древнескандинавским считаем совокупность всех скандинавских диалектов в период с вырождения протоскандинавского языка-основы (urnordisk) до общепринятой точки

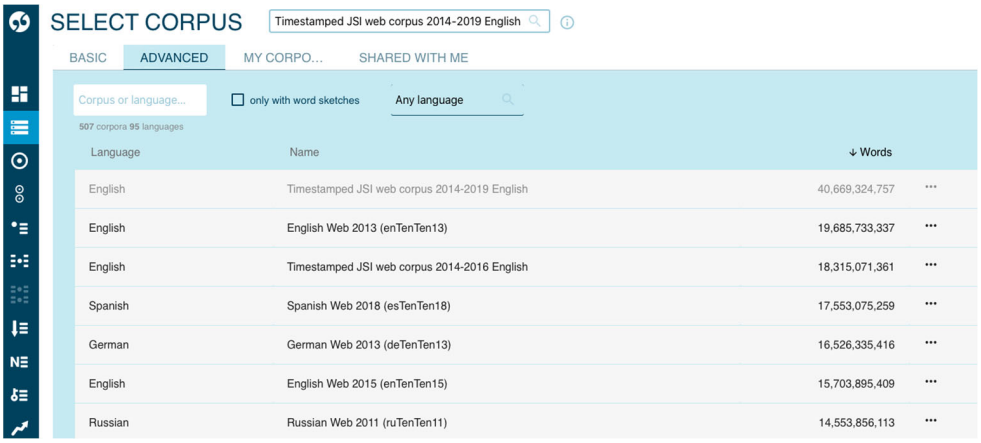

Рис. 1. Корпуса SketchEngine с сортировкой по объему

Fig. 1. SketchEngine corpora, sorted by size

распада на отдельные языки, т. е. с IX по XIII в., за исключением гутнийского, фонетика которого наследует многие признаки протогерманского языка.

В качестве корпусного материала взят текст норвежского образовательного трактата, датируемого примерно 1250 г., «Царское зерцало» (Konungs skuggsjá), написанного в целях воспитания короля Магнуса Лагабёте в виде диалога с его отцом Хоконом Хоконссоном [6]. Текст объемом около 60 тыс. слов-токенов взят из корпуса Menota, размещенного в открытом доступе на платформе Clarino [12], содержит полную частеречную разметку и приводится в дипломатической записи. В целях упрощения анализа используемый набор частеречных помет сокращен со 100 до 9, т. е. итоговый набор включает 9 классов: существительные, глаголы, союзы, детерминативы, наречия, предлоги, местоимения, прилагательные и причастия, распределившиеся следующим образом (см. рис. 2).

Текст был преобразован в векторный формат методом скользящего среднего, рекомендуемым Пири Такалой применимо к морфологически сложным языкам [17]; метод генерирует репрезентацию слова w = ($w_a w_b ... w_z$), где

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

где c — индекс обрабатываемого символа (0 для первого символа в слове, 1 для второго и т. д.), α соответствует гиперпараметру, обусловливающему понижение значения на выходе, Z — нормализатор, значение которого пропорционально длине слова. Размерность получаемых на выходе векторов, каждый из которых соответствует одному слову исходного текста, равно числу символов алфавита, умноженному на три (по одному измерению на символ при расчете по указанной

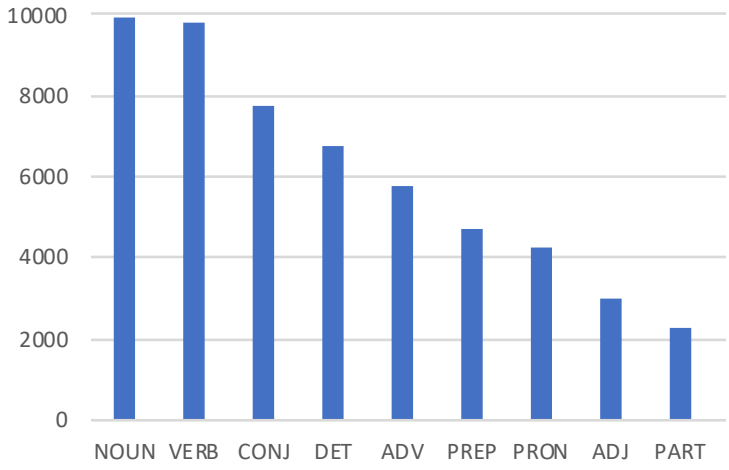

Рис. 2. Распределение частей речи в тексте «Царского зерцала» (Konungs skuggsjá, Konungaspeilet)

Fig. 2. PoS distribution of *King's Mirror* (Konungs skuggsjá, Konungaspeilet)

формуле, который повторяется в обратном порядке, т. е. *слово — оволс*, а также по одному измерению на символ с указанием того, сколько раз он встречается в слове). Алфавит анализируемого текста состоит из 47 символов, т. е. размерность векторов составит 47*3 = 141.

Далее полученная выборка (около 60 тыс. 141-мерных векторов с присвоением одного из 9 классов каждому из них) использовалась для обучения алгоритма-ансамбля, известного как модель случайного леса (англ. RFM, random forest model), дополнительно усиленного методом AdaBoost [18]; качество обучения проверялось методом 10-проходной перекрестной проверки. Полученные в итоге результаты классификации и отдельные примеры ошибок проанализированы с позиции исторических процессов и особенностей, отличающих древнескандинавский язык: фонетических изменений и морфограмматических качеств, имеющих то или иное орфографическое проявление.

Результаты и обсуждение

По итогам классификации текстов получены результаты, приведенные в таблице 1.

Точность, полнота и F-показатель классификации составляют 97,32%, что превышает результаты, достигнутые ранее в исследовании частеречного классификатора на основе TreeTagger, предварительно обученного на материале современного исландского языка, применимо к древнеисландскому — 90% [11]. В этой связи представляется целесообразным проанализировать те немногочисленные ошибки, которые были допущены классификатором, см. таблицу 2.

Очевидны следующие паттерны коллизии классов:

Таблица 1

Результативность классификатора RFM + AdaBoost

Table 1

RFM + AdaBoost classification performance

Класс	Точность	Полнота	F-показатель
NOUN	0,921	0,967	0,943
DET	0,995	0,994	0,994
PRON	0,999	0,996	0,998
VERB	0,961	0,959	0,960
ADV	0,991	0,983	0,987
PREP	0,999	0,997	0,998
CONJ	0,999	0,999	0,999
PART	1,000	1,000	1,000
ADJ	0,940	0,818	0,875
Средневзвешенное значение	0,973	0,973	0,973

— **существительное**: определяется почти всегда корректно, коллидирует с глаголом и прилагательным;
— **детерминативы** и **местоимения**: коллизия отсутствует (здесь и далее порогом коллизии считаем 1% от общего числа примеров той или иной части речи);
— **глагол**: коллизия с существительным и прилагательным;
— **наречие**, **предлог**, **союз**, **причастие**: коллизия отсутствует;
— **прилагательное**: коллизия с существительным, глаголом и наречием.

Таким образом, (практически) 100%-я точность и полнота классификации наблюдаются в отношении тех частей речи, которые не изменяются по форме (наречие, служебные части речи) и/или представлены весьма ограниченным набором (детерминативы и местоимения); интересно, что, несмотря на морфологическую схожесть, причастие не коллидирует с прилагательным. Значительная перекрестная коллизия наблюдается в отношении трех самостоятельных частей речи: имен существительного и прилагательного, а также глагола. Далее рассмотрим причины подобного явления.

В рамках анализа ошибок классификации необходимо обратиться к принципам работы самого классифицирующего алгоритма. В основе применяемой методики лежала модель случайного леса, генерирующая деревья принятия решений, вывод которых при обработке представленных в векторном виде данных зависит от отдельных, значимых признаков в имеющемся векторном пространстве. Каждый признак в использовавшейся репрезентации соответствовал одному из символов; так как оба алгоритма исполнялись в среде Weka [20], было принято решение оценить значимость признаков с помощью встроенной функции ClassifierAttribute-

Таблица 2
Матрица ошибок классификатора

Table 2
Classifier confusion matrix

Классиф. как	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**
a = NOUN	9 265	14	1	224	7	0	3	0	70
b = DET	31	6 543	0	12	4	0	0	0	4
c = PRON	8	1	4 071	3	2	0	2	0	1
d = VERB	335	11	0	9 062	10	2	2	0	50
e = ADV	54	7	1	16	5 445	0	1	0	29
f = PREP	9	0	0	3	3	4 555	0	0	0
g = CONJ	2	0	0	4	1	0	7 583	0	0
h = PART	0	0	0	0	0	0	0	2 160	0
i = ADJ	384	5	0	126	17	0	0	0	2 404

Eval. Наиболее значимыми оказались атрибуты, соответствующие следующим символам (в порядке убывания значимости): *a, i, ð, e, n, o, t, r, u*.

Использовавшийся в рамках исследования текст был обработан с помощью алгоритма Morfessor, позволяющего без предварительного обучения идентифицировать морфемы в тексте с помощью скрытых марковских моделей [14]. Данный алгоритм идентифицировал в тексте следующие частотные морфы, перечисленные в порядке убывания частотности (указана в скобках, описание дано преимущественно по [7], также по [15], примеры — непосредственно из исследуемого текста):

- **-r** (516). Применяется в качестве личного окончания глаголов в настоящем времени и ед. ч.: *fundr sem byria**r*** («битва, что начинается»); в качестве маркера мужского рода и именительного падежа прилагательных: *dyrlig**r** keisari* («дражайший кайзер»); аналогично у существительных: *guð sá at engi mað**r** vita* («бог узрел, что никто из людей не знал»); также в качестве маркера мн. ч. местоимений: *mini**r** mæn* («мои люди»).
- **-a, -a-** (480). Является окончанием преимущественно мн. ч. в косвенных падежах у прилагательных, существительных и местоимений, от которых происходит адъективальная парадигма (личных и указательных): *all**a**, mann**a**, min**a***; инфиксом у глаголов, в том числе предшествующим временному маркеру, как в *loy**a**ði*, или служащим для соединения двух семантических основ в составных словах: *rað**a**gera* («давать совет»). Последнее утверждение, однако, требует некоторого дополнительного разъяснения: в подобных случаях речь может идти о составлении слов по родительному падежу (исл. eignarfallsamsetning, англ. genitive-case composition). В глаголах также маркирует инфинитив и/или мн. ч.: *mænn þycki**a*** («люди презирают»).
- -**(u)m** (342). Маркер датива мн. ч. существительных: *augum* («глазам»), прилагательных: *baðu**m*** («обоим»), а также личных местоимений: *minu**m*** («моим»); также указывает на мн. ч. первого лица глаголов: *eru**m*** («есмы»).
- -**s** (311). Маркер родительного падежа существительных мужского и среднего рода в **a**-склонении, отчасти в **ia**-склонении: *dag**s*** («дня»), также у прилагательных: *mykil**s*** («многого»).
- -**ar** (302). Маркер родительного падежа и мн. ч. в прочих склонениях: *drotning**ar***, также у прилагательных *æng**ar***, *mikill**ar***.
- -**t** (290). Дентальный суффикс причастия глаголов: *ek hæfe spur**t*** («я спросил»), встречается в некоторых наречиях: *sam**t*** («все равно»), маркирует средний род прилагательных сильного склонения: *stor**t*** («большое»).
- -**er** (187). См. -**ir**. Также входит в основу некоторых существительных, частотных в данном тексте: *faðe**r*** («отец»).
- -**u** (178). Суффикс, преимущественно свойственный прилагательным и существительным в отдельных неноминативных формах, преимуществен-

но в винительном падеже: *aug**u***, *bað**u***. Также ложно определяется в клитикализованном императиве глаголов: *gerð**u*** («сделай»).

- **-ir** (151). Суффикс мн. ч. прилагательных: *aðr**ir*** («другие») и ед. ч. глагола, преим. 2 лица: *kaup**ir*** («покупаешь»).
- **-i, -i-** (96). Преимущественно реализуется как маркер датива ед. ч. существительных: *konong**i*** («королю»), в том числе в виде инфикса, предшествующего маркеру определенности: *fisk**i**num* («рыбе», определенная форма, фактически речь идет об агглютинации).
- **-an** (91). Окончание аккузатива сильного склонения прилагательных в мужском роде: *all**an*** («всех»), *stor**an*** («больших»). Встречается у наречий, где не является самостоятельным морфом, а входит в основу: *sam**an*** («вместе»). Неоднократно наблюдается у служебных слов, образовавшихся в результате их слияния: *ut**an***, *sið**an***, однако коллизия с ними незначительна или отсутствует.
- **-inn** (71). Маркер определенности номинатива ед. ч. мужского рода: *drott**inn*** («владыка»).
- **-liga** (71). Суффикс наречий: *rang**liga*** («неверно»), также прилагательных: *ýmis**liga*** («различные»), причем в последних является комбинацией адъективального маркера и окончания -a из общеименной парадигмы.
- **-na** (71). Маркер аккузатива определенной формы существительных женского рода в ед. ч., мужского и среднего рода в мн. ч.: *kona**na*** («женщину»); встречается в наречиях: *ger**na*** («с радостью»), *sam**na*** («вместе»); в глагольном причастии: *gef**na*** («данная») от *gefa* (второе причастие в данном случае наследует парадигму прилагательных, причем суффикс -n отмечается только у сильноглагольных причастий, т. к. слабоглагольные используют дентальную суффиксацию).
- **-ðe** (61). Маркер претеритума глаголов: *haf**ðe*** («имел»); также нередко встречается у слов других ЧР, где ð является частью корня: *bæ**ðe***, «оба».
- **-ra** (59). В основном употребляется в качестве суффикса прилагательных в положительной и сравнительной степенях в сильном склонении, причем в обоих случаях является составным, например, в *lang**ra*** («длинных»), где маркер -r указывает на родительный падеж, а -a — на мн. число. Также идентифицирован в многочисленных глаголах, таких как *hey**ra*** («слышать») или *læ**ra*** («учить»), где не является морфом как таковым, т. к. в них -r относится к корню, а -a — маркер инфинитива или мн. числа. Тем не менее коллизии глагола и прилагательного в подобных случаях не возникает.
- **-ði** (57). Пример: *ælska**ði*** («любил»). См. **-ðe**.
- **-num** (51). Маркер датива множественного числа определенной формы существительных независимо от рода: *dægi**num*** («дням» в определенной форме, обр. внимание на i-мутацию в корне).

В свете вышеописанных морфологических особенностей рассмотрим наиболее показательные примеры возникшей коллизии, представленные в таблице 3.

В то время как морф *-a* представляется проблематичным, так как многократно определен алгоритмом в тех случаях, когда таковым не является (в особенности когда выявлен в качестве префикса или инфикса, например, *agnum* или *saum*, причем употребления в виде «префикса» исключены из приведенной выше статистики), в целом выявление морфов можно охарактеризовать как точное. Отметим, что символы, которые ранее были определены как наиболее значимые при частеречной разметке векторизованного скользящим средним текста методом случайного леса, фактически являются теми символами, из которых состоят наиболее частотные идентифицированные алгоритмом Morfessor морфы. Таким образом можно заключить, что выбранный метод векторизации обеспечивает эффективное кодирование морфологических признаков слова и обращение алгоритма-классификатора к ним.

Зачастую основным определяющим признаком частеречной принадлежности оказывается суффигированный маркер определенности, например, *-inn* в мужском роде. Происхождение этого маркера до конца не установлено, однако наиболее современная точка зрения заключается в том, что он возник ввиду клити-

Таблица 3

Примеры коллизии

Table 3

Collision examples

Слово	Фактическая ЧР	Классиф. как	Комментарий
tænr	Существительное	Глагол	-r
klæðe	Существительное	Глагол	-ðe является частью корня
prætta	Существительное	Глагол	-a; схоже с глаголом *þrætta*
þægna	Существительное	Глагол	-na
spurnum	Глагол	Сущ.	-m
bragðar	Глагол	Сущ.	-r
mantu	Глагол	Сущ.	-tu является маркером императива ед. ч., но -u наиболее часто встречается в существительных
kunner	Прилагательное	Глагол	-r
utrulegr	Прилагательное	Сущ.	-r
náliga	Прилагательное	Наречие	-liga
sunnaʀr	Наречие	Сущ.	-r как маркер сравнительной степени; в целом слово схоже с son(n)r, «сын»
ínnan	Наречие	Сущ.	-an
bœtr	Наречие	Глагол	-r

кализации указательного местоимения *hinn* и ему подобных (например, *hið* в среднем роде): *dagr hinn — dagrinn*; *dags hins — dagsins*. Данное обстоятельство приводит к коллизии существительного среднего рода в определенной форме с глаголом, чьи причастные формы могут оканчиваться на *-ð* — традиционный маркер прошедшего времени и прошедшего причастия в германских языках [16]. Так, алгоритм определил слово *astsæð* как глагол, хотя оно является именно существительным.

Маркер *-r* в окончаниях практически во всех случаях является продуктом ротацизма, т. е. превращения протогерманского *z* сначала в *ʁ*, затем в *r*, (*dagr* происходит от протогерм. **dagaz*), причем на момент написания исследуемого текста слияние двух ротических звуков еще было неполным, что отражено на письме, например, в виде чередования *faðer/faðeʁ*. Данное обстоятельство относится как к существительным, так и к прилагательным: ср. *goðr* и **gōdaz*, «хороший»; и к глаголам: *heldr*, **haldizi* («держишь»; обратим внимание на i-мутацию корневой гласной). Таким образом, омоморфия, наблюдаемая в исследуемом тексте, фактически восходит к протогерманскому языку.

Отметим, что за счет полного ротацизма протогерманского /z/ коллизии не вызывает маркер генитива мужского и среднего рода *-s*, который, к примеру, присутствует в глагольной парадигме древнеанглийского языка из-за возникновения в нем /s/ из протогерманского /z/ по вернеровскому процессу: *wæs, wæron* («был», «были») [19].

Общим для глаголов и именных частей речи является суффикс *-um*, при этом источник у него практически идентичен: *m* как маркер датива у существительных и мн. ч. первого лица глаголов имелся еще в праиндоевропейском языке и сохранился до сих пор во многих ИЕЯ, ср. протогерм. **daga**__**m**__*az* («дня**м**») и **halda**__**m**__*az* («держи**м**»). Аналогичный суффикс проявляется в определенных формах в связи с их происхождением как клитики указательного местоимения, имевшего схожую парадигму. Клитика наблюдается в императиве, см. пример в таблице: *mantu — mana þu* («помни»).

Интересно, что несмотря на наличие формообразовательной функции умлаута в древнескандинавском языке корневые гласные не были идентифицированы алгоритмом Morfessor, что, на наш взгляд, обусловлено, с одной стороны, их нахождением в центральной позиции в основе, с другой — малой репликативностью.

Заключение

В качества общего вывода по выполненному исследованию можно отметить, что в то время как выбранный метод кодирования действительно обеспечивает репрезентацию морфологических маркеров в тексте, на что указывает совпадение символьных признаков с высоким весовым коэффициентом и выделенных сторонним алгоритмом частотных морфов, самой по себе такой репрезентации недостаточно, так как многие части речи обладают идентичными морфографемами, унаследованными из протогерманского языка, а в некоторых случаях к омоморфии приводят клитические процессы или даже наличие той или иной

графемы в конце корня слова. При этом часть слов, классифицированных неверно, могли бы быть классифицированы корректно за счет применения алгоритмов, основанных на пространственном (последовательном) распределении, то есть обращающихся к контексту, таких как скрытые марковские модели, что задает перспективное направление дальнейшей исследовательской работы.

СПИСОК ЛИТЕРАТУРЫ

1. Арапов М. А. Математические методы в исторической лингвистике / А. М. Арапов, М. М. Херц. М.: Наука, 1973. 322 с.
2. Николаева Н. А. Тематизация презенса сильного глагола в кельтских и германских языках: дис. … канд. филол. наук / Н. А. Николаева. М.: МГУ им. Ломоносова, 2003. 200 с.
3. Bandle O. The Nordic languages: an international handbook of the history of the North Germanic languages / O. Bandle, K. Braunmüller, E. H. Jahr, A. Karker, H. P. Naumann, U. Telemann, L. Elmevik, G. Wildmark (eds.). Berlin: De Gruyter Mouton, 2002. 1084 p.
4. Gade K. E. Homosexuality and rape of males in Old Norse law and literature / K. E. Gade // Scandinavian Studies. 1986. Vol. 58. No 2. Pp. 124-141.
5. Hagland J. R. A note on Old Norwegian vowel harmony / J. R. Hagland // Nordic Journal of Linguistics. 1978. Vol. 1. Pp. 141-147.
6. Haugen O. E. Norrøne tekster i utval / O. E. Haugen. Oslo: Ad Notam Gyldendal, 1994. 312 p.
7. Haugen O. E. Grunnbok i norrønt språk / O. E. Haugen. Oslo: Ad Notam Gyldendal, 1995. 320 p.
8. Jahr E. H. Historisk språkvitenskap / E. H. Jahr, O. Lorentz. Oslo: Novus, 1993. 431 p.
9. Karttunen L. Applications of finite-state transducers in natural language processing / L. Karttunen // Implementation and Application of Automata, 5th International Conference, CIAA 2000 (July 24-25, 2000). Pp. 34-46.
10. Kytö M. Manual to the Diachronic Part of the Helsinki Corpus of English Texts / M. Kytö. Helsinki: University of Helsinki, 1996. URL: http://clu.uni.no/icame/manuals/HC/INDEX.HTM
11. Loftsson H. Improving the PoS tagging accuracy of Icelandic text / H. Loftsson, I. Kramarczyk, S. Helgadóttir, E. I. Rögnvaldsson // Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009). Odense, Denmark: Northern European Association for Language Technology (NEALT), 2009. Pp. 103-110.
12. Medieval Nordic Text Archive. URL: http://clarino.uib.no/menota/page (дата обращения: 05.11.2019).
13. Silva A. P. An approach to the POS tagging problem using genetic algorithms / A. P. Silva, A. Silva, I. Rodrigues // Computational Intelligence. Berlin: Springer, 2015. Pp. 3-17.
14. Smit P. Morfessor 2.0: toolkit for statistical morphological segmentation / P. Smit, S. Virpioja, S. A. Grönroos, M. Kurimo // 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.

15. Spurkland T. Innføring i norrønt språk / T. Spurkland. Oslo: Universitetsforlaget, 1989. 173 p.
16. Stroh-Wollin U. The emergence of definiteness marking in Scandinavian — new answers to old questions / U. Stroh-Wollin // Arkiv för nordisk filologi. 2016. No 131. Pp. 129-169.
17. Takala P. Word embeddings for morphologically rich languages / P. Takala // European Symposium on Artificial Neural Networks (Bruges, April 27-29, 2016). Pp. 177-182.
18. Tharwat A. AdaBoost Classifier: An Overview / A. Tharwat. Frankfurt: Frankfurt University of Applied Sciences, 2018. URL: https://www.researchgate.net/publication/323119678_AdaBoost_classifier_an_overview
19. Vrieland S. D. Old English and Old Norse. An Introduction to West and North Germanic / S. D. Vrieland. Copenhagen: University of Copenhagen, 2004. URL: https://www.academia.edu/7017345/Old_English_and_Old_Norse_Introduction_to_West_and_North_Germanic
20. Witten H. I. Data Mining: Practical Machine Learning Tools and Techniques / H. I. Witten, E. Frank, M. A. Hall. Burlington, Massachusetts: Morgan Kaufmann Publishers Inc., 2011. 664 p.

Raul D. KARIMOV[1]

UDC 81-112

OLD NORSE AS A PROBLEM OF AUTOMATIC POS-TAGGING

[1] Postgraduate Student, Department of Romance and Germanic Languages and Cross-Cultural Communication,
Chelyabinsk State University
raoul.karimov@hotmail.com; ORCID: 0000-0003-0313-0309

Abstract

This article dwells upon automatic PoS-tagging of Old Norse by computational means, including machine learning. It analyzes the available language material in diachrony from the standpoint of how language evolution might have affected the quality of automatic PoS-tagging. This article further describes the phonetic traits that have assumingly led to any classification errors.

The research material is an Old Norwegian educational text titled *Konungs skuggsjá*, or "King's Mirror", vectorized by the moving average method and then used to train an Ada-Boosted random forest model. The resulting classification accuracy is about 97%. However, being non-contextual, this vectorization method enables no complete differentiation of morphologically similar parts of speech: verbs, nouns, adjectives, and adverbs. This becomes evident when digging into the identified high-weight classification features, each being a vectoral dimension corresponding to a specific alphabet character; another indicative factor comprises Morfessor-identified high-rank morphs, analyzing which reveals the morphogrammatic units that cause the most classification errors.

Historical consideration of these morphs shows that their collision is due to them being inherited from Proto-Germanic (PG) while undergoing rhotacism, or conversion from PG /z/ to ON /r/. However, the same process effectively prevents the collision of rhotacized finite verbal forms with the genitive case that inherits the PG suffix *-s*.

The key finding is that such morphological collision being unavoidable, character-based vectorization might not suffice when using a small training set or when trying to classify not only by parts of speech, but also by specific forms in the paradigm.

Citation: Karimov R. D. 2019. "Old Norse as a Problem of Automatic PoS-tagging". Tyumen State University Herald. Humanities Research. Humanitates, vol. 5, no 4 (20), pp. 38-52.
DOI: 10.21684/2411-197X-2019-5-4-38-52

Keywords

Old Norse, corpus linguistics, phonetics, morphology, parts of speech, annotation, tagging, vector representation.

DOI: 10.21684/2411-197X-2019-5-4-38-52

REFERENCES

1. Arapov M. A., Hertz, M. M. 1973. Mathematical Methods in Historical Linguistics. Moscow: Nauka. [In Russian]
2. Nikolayeva N. A. 2003. "Thematization of the present tense of strong verbs in celtic and germanic languages". Cand. Sci. (Philol.) diss. Moscow: Moscow State University. [In Russian]
3. Bandle O. 2002. The Nordic languages: an international handbook of the history of the North Germanic languages. Edited by O. Bandle, K. Braunmüller, E. H. Jahr, A. Karker, H. P. Naumann, U. Telemann, L. Elmevik, and G. Wildmark. Berlin: De Gruyter Mouton.
4. Gade K. E. 1986. "Homosexuality and rape of males in Old Norse Law and literature". Scandinavian Studies, vol. 58, no 2, pp. 124-141.
5. Hagland J. R. 1978. "A note on Old Norwegian vowel harmony". Nordic Journal of Linguistics, vol. 1, pp. 141-147.
6. Haugen O. E. 1994. Norrøne tekster i utval. Oslo: Ad Notam Gyldendal. [In Norwegian]
7. Haugen O. E. 1995. Grunnbok i norrønt språk. Oslo: Ad Notam Gyldendal. [In Norwegian]
8. Jahr E. H., Lorentz O. (eds.). 1993. Historisk språkvitenskap. Oslo: Novus. [In Norwegian]
9. Karttunen L. 2000. "Applications of Finite-State Transducers In Natural Language Processing". Proceedings of the 5th International Conference "Implementation and Application of Automata", CIAA 2000 (24-25 July), pp. 34-46.
10. Kytö M. 1996. Manual to the Diachronic Part of the Helsinki Corpus of English Texts. Helsinki: University of Helsinki.
11. Loftsson H., Kramarczyk I., Helgadóttir S., Rögnvaldsson E. I. 2009. "Improving the PoS tagging accuracy of Icelandic text". Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009), pp. 103-110. Odense, Denmark: Northern European Association for Language Technology (NEALT).
12. Medieval Nordic Text Archive. Accessed 11 May 2019. http://www.clarino.uib.no/menota
13. Silva A. P., Silva A., Rodrigues I. 2015. "An approach to the POS tagging problem using genetic algorithms". In: Computational Intelligence, pp. 3-17. Berlin: Springer.
14. Smit P., Virpioja S., Grönroos S.A., Kurimo M. 2014. "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, pp. 21-24. Stroudsburg, PA, USA: Association for Computational Linguistics.
15. Spurkland T. 1989. Innføring i norrønt språk. Oslo: Universitetsforlaget. [In Norwegian]
16. Stroh-Wollin U. 2016. "The emergence of definiteness marking in Scandinavian — new answers to old questions". Arkiv för nordisk filologi. 2016, no 131, pp. 129-169.

17. Takala P. 2016. “Word embeddings for morphologically rich languages”. European Symposium on Artificial Neural Networks (27-29 April, Bruges, Belgium), pp. 177-182.
18. Tharwat A. 2018. AdaBoost Classifier: An Overview. Frankfurt: Frankfurt University of Applied Sciences.
19. Vrieland S. D. 2004. Old English and Old Norse. An Introduction to West and North Germanic. Copenhagen: University of Copenhagen.
20. Witten H. I. 2011. Data Mining: Practical Machine Learning Tools and Techniques. Burlington, Massachusets: Morgan Kaufmann Publishers Inc.