Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт (государственный университет)»

Факультет инноваций и высоких технологий
Кафедра анализа данных

Направление подготовки: 01.03.02 Прикладная математика и информатика

Автоматическая расстановка ударений в словах

Бакалаврская работа

Обучающийся: ФИО

Научный руководитель: должность
ФИО

Москва 2018

Аннотация

Умные слова в 1500 знаков

Оглавление

Введение

Ударение в словах – важнейший элемент устной, письменной и внутренней речи. В русском языке оно играет исключительно важную роль, так как благодаря ему мы можем различать слова. Одной из сложностей русского языка является его свободное ударение, которое не закреплено за каким-либо определенным слогом или морфемой слова. Любой слог может выделяться фонетически. К тому же ударение может меняться с изменением грамматической формы слова. Как отмечает лингвист Н. А. Еськова, «слова с подвижным ударением в русском языке исчисляются сотнями. В процентном отношении это немного, но среди них много чрезвычайно употребительных, поэтому в речи они достаточно заметны» [1]. Например: фла́г — фла́га — фла́ги; но вра́г — врага́ — враги́

Есть языки, где ударение всегда на одном и том же слоге — такое ударение называют фиксированным. Например, во французском ударение всегда на последнем слоге, в польском — на предпоследнем, в чешском — на первом. В русском языке аналогичные правила весьма размыты, поэтому если человек не знает, как правильно ставить ударение в слове, то по одному только его внешнему облику сделать правильный выбор бывает сложно. Нет общих правил ударения и в заимствованных для русского языка словах . Иногда оно меняет свое место по сравнению с ударением в языке-источнике: ноутбу́к, скелето́н, футбо́л, хокке́й. А иногда сохраняет: бульо́н, гардеро́б, жалюзи́. Расстановка ударений, как часть задачи предсказания произношения, - важная составляющая приложений, таких как: автоматическое распознавание речи, синтез речи, транслитерация. Кроме того – это необходимо всем, изучающим русский язык.

Глава 1.
Литературный обзор

Работы по предсказанию постановки ударений в словах велись в двух направлениях. На основе лингвистических правил [2, 3] и на основе анализа данных, где модели строятся напрямую из текстов с обозначенными ударениями.

В русском языке сохраняется множество индо-европейских шаблонов ударений. Чтобы узнать ударение морфологически сложного слова, состоящего из основы и окончания, необходимо узнать является ли основа ударной и на какой слог падает в ней ударение, либо ударным является окончание [4].

1.1. Предсказание ударения в слове на основе ранжирования

Авторы [5] рассмотрели проблему расстановки ударений, как задачу ранжирования. В своем исследовании они опирались на более раннюю статью [6]. Из каждого слова выделяются гласные буквы и они предполагаются, как возможные варианты постановки ударений. Целью модели является отранжировать варианты так, чтобы верная гипотеза имела наименьший ранг.

Для ранжирования гипотез применялось Maximum Entropy ранжирование [7]. Во время обучения модели ей подавался набор правильных гипотез и их признаков. Во время предсказания в модель подавались все гипотезы, и в качестве верной выбиралась гипотеза с максимальным предсказанным результатом. В качестве основы для ранжирования использовалась линейная модель, вместо SVM, так как она более эффективна с вычислительной точки зрения для обучения и применения.

В базовой статье [6] признаками являлись триграммы для гласных букв следующего вида: предыдущая согласная, если она есть, гласная буква, следующая за ней слогласная, если она есть (Dou). На основе лигвистического

исследования в данной статье авторы добавили следующие признаки: для каждого слова взяты все начальные и конечные части (уже - у, уж, уже, е, же) (Affix). Также эти признаки добавлены в следующем виде: все буквы заменены на их абстрактные фонетические классы (представлены в табл. 1.1)(Abstr Aff).

Таблица 1.1 — Абстрактные фонетические классы

Класс	Буквы
vowel	а, е, и, о , у, э, ю, я, ы
stop	б, д, г, п, т, к
nasal	м, н
fricative	ф, с, ш, щ, х, з, ж
hard/soft	ъ, ь
yo	ё
semivowel	й, в
liquid	р, л
affricate	ц, ч

В качетсве данных авторы использовали Грамматический словарь русского языка Зализняка[8], разбитый на обучающую и тестовую выборки. Из тестовой выборки также были отдельно выделены те слова, которые не встречались в обучающей выборке, и для них также были получены результаты. Результаты экспериментов представлены в табл. 1.2.

Таблица 1.2 — Результаты ранжирования

Признаки	Accuracy score
Тестовая выборка	
Dou	0.972
Aff	0.987
Aff+Abstr Aff	0.987
Dou et al+Aff	0.987
Dou et al+Aff+Abstr Aff	0.987
Слова не встречавшиеся в обучающей выборке	
Dou	0.806
Aff	0.798
Aff+Abstr Aff	0.810
Dou et al+Aff	0.823
Dou et al+Aff+Abstr Aff	0.89

Таблица показывает влияние взаимодействия признаков на обобщающую способность модели, и лучший результат достигнут при использовании всех признаков.

Недостатками этой статьи является не использование контекста для определения места ударения. При подсчете результатов никак не учитывалась частота употребления слов в текстах языка, использовался просто его лексический набор.

1.2. Предсказание ударения в слове при помощи конечного преобразователя

Целью авторов[9] было разработки модели, которая могла быть помочь людям изучать русский язык, они решили что в некоторых словах ударение может быть пропущено. Авторы считали, что неправильное ударение может быть хуже, чем его отсутствие для человека осваивающего новый язык.

Модель состояла из двух частей: конечного преобразователя [10, 11], который из полученного слова генерировал все возможные, корректные по его мнению, позиции ударения. Далее при помощи формальной грамматики [12] удалялись варианты, которые не подходили по контексту. Если после применения этой процедуры оставался один вариант прочтения, то он и

выбирался как финальный. Если ни одного – то ударение в слове не проставлялось. Если же вариантов было несколько – то в зависимости от эксперимента выбиралась дальнейшая стратегия.

Авторами использовался корпус текстов, состоящий из 7689 слов с размеченными ударениями, это были тексты для начинающих изучать русский язык. Также для обучения модели применялся Грамматический словарь русского языка Зализняка [8].

Описание экспериментов:

- **bare:** при нескольких возможных вариантах прочтения слова, ударение в слове не проставлялось.
- **safe:** при нескольких возможных вариантах прочтения, ударение в слове выставлялось, если во всех них ударение падало на один и тот же слог.
- **randReading:** при нескольких возможных прочтениях, случайно выбиралось одно с вероятностью выбора варианта равной частоте встречаемости этого варианта в тексте.
- **freqReading:** при нескольких возможных прочтениях, выбирается вариант с максимальной частотой встречаемости среди всех вариантов в тексте.

Эксперименты были проведены при использовании формальной грамматики с учетом контекста и без него. Результаты представлены в табл. 1.3. Для слов, которые не встретились в словаре, применялось простое правило постановки ударения: ударение падает на последню гласную, после которой идет согласная. Это является наиболее вероятным вариантом ударения в русском языке [13]. В результатх это отображено как guessSyl.

Таблица 1.3 — Результаты применения конечного преобразователя

Эксперимент	Accuracy score	Доля ошибок	Доля пропущенных слов
Без грамматики			
bare	30.43	0.17	69.39
safe	90.07	0.49	9.44
randReading	94.34	3.36	2.30
freqReading	95.53	2.59	1.88
randReading+guessSyll	94.99	4.05	0.96
freqReading+guessSyll	95.83	3.46	0.72
С грамматикой			
bare	45.78	0.44	53.78
safe	93.21	0.74	6.058
randReading	95.50	2.59	1.90
freqReading	95.73	2.40	1.88
randReading+guessSyll	95.92	3.33	0.74
freqReading+guessSyll	96.15	3.14	0.72

При использовании модели без формальной грамматики полнота гипотез составила 97.55%, что является максимумом результата для данной модели. При использовании грамматики полнота составила 97.35%. Эти результаты являются потолком для соответствующих экспериментов. Совмещение всех моделей и предсказываение методом FreqReading позволило получить наибольших процент правильных ударений. Метод расстановки ударений для неизвестных слов в данном случае имеет точность всего 21%. При этом высокая точность была достигнута за счет расстановки ударений почти во всех словах, что соответственно повысило уровень ошибок.

Метод, представленный в этой статье, является попыткой улучшить словарный метод, путем разрешения неоднозначностей в омографах при помощи формальной грамматики. При этом для слов, которые не встретились в словаре работает очень простой и слабый алгоритм. В этом случае качество получается очень низким. Недостатком является также использование небольшого закрытого корпуса текстов. А так как использовались тексты для начинающих изучать язык, их словарь скорее всего был достаточно мал. Не ясна цель проведения эксперимента RandReading, так как несложно показать

строго математически, что метод FreqReading всегда дает большую вероятность правильного ответа.

1.3. Предсказание ударения в слове при помощи символьной нейронной сети

В качестве основы для модели авторы [14] использовалась символьная двусторонняя рекуррентная нейронная сеть на основе LSTM-модулей. На вход подавалась матрица размера [длина фразы; число возможных символов]. Для кодирования символов было применено one-hot кодирование. Авторами выбрана следующая архитектура: к входной матрице применяется двусторонняя рекуррентная нейронная сеть, сконкатенированные вектора, полученные от рекуррентного слоя подаются в полносвязные слой с softmax активацией. На выходе получается вектор размера равного длине фразы, соответствующий распределению вероятностей постановки ударения в конкретной позиции.(представлена на рис. 1.1)

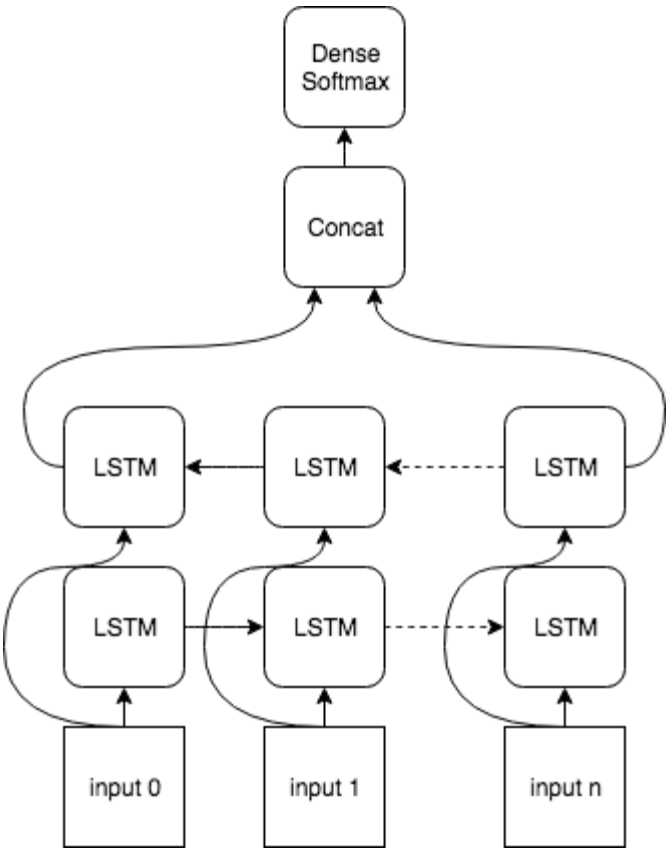

Рисунок 1.1 — Архитектура сети

Для разных экспериментов использовался грамматический словарь русского

языка Зализняка [8] и база данных акцентологической разметки в составе национального корпуса русского языка [15].

Авторами были проведены следующие эксперименты

1. **Обучение и предсказание на основе словаря.** Словарь Зализняка был разделен на обучающую и тестовую выборки в соотношении 2:1. Результаты этого эксперимента представлены в табл. 1.4.

2. **Обучение и предсказание на основе акцентологического корпуса.** С корпусом было проведено два эксперимента, в первом в качестве фразы использовалось только само слово. Во втором же, к нему были дописаны три последние буквы из слова, которое идет перед ним в предложении, если такое было. Сравнительные результаты экспериментов представлены в табл. 1.5. Основная разница между моделями с контекстом и без него может быть видна только на омографах. Результаты применения на них представлены в табл. 1.6. Как видно из результатов модель успешно использует контекст для расстовки ударения в омографах во многих случаях.

Таблица 1.4 — Результаты применения нейросетевой модели на словаре Зализняка

Эксперииимент	Accuracy score	Доля Ошибок	Доля пропущенных слов
Без грамматики			
bare	30.43	0.17	69.39
safe	90.07	0.49	9.44
randReading	94.34	3.36	2.30
freqReading	95.53	2.59	1.88
randReading+guessSyll	94.99	4.05	0.96
freqReading+guessSyll	95.83	3.46	0.72
С грамматикой			
bare	45.78	0.44	53.78
safe	93.21	0.74	6.058
randReading	95.50	2.59	1.90
freqReading	95.73	2.40	1.88
randReading+guessSyll	95.92	3.33	0.74
freqReading+guessSyll	96.15	3.14	0.72

Таблица 1.5 — Результаты применения нейросетевой модели на словаре Зализняка

Эксперимент	Accuracy score	Доля Ошибок	Доля пропущенных слов
Без грамматики			
bare	30.43	0.17	69.39
safe	90.07	0.49	9.44
randReading	94.34	3.36	2.30
freqReading	95.53	2.59	1.88
randReading+guessSyll	94.99	4.05	0.96
freqReading+guessSyll	95.83	3.46	0.72
С грамматикой			
bare	45.78	0.44	53.78
safe	93.21	0.74	6.058
randReading	95.50	2.59	1.90
freqReading	95.73	2.40	1.88
randReading+guessSyll	95.92	3.33	0.74
freqReading+guessSyll	96.15	3.14	0.72

Таблица 1.6 — Результаты применения нейросетевой модели на омографах

Эксперимент	Accuracy score	Доля Ошибок	Доля пропущенных слов
Без грамматики			
bare	30.43	0.17	69.39
safe	90.07	0.49	9.44
randReading	94.34	3.36	2.30
freqReading	95.53	2.59	1.88
randReading+guessSyll	94.99	4.05	0.96
freqReading+guessSyll	95.83	3.46	0.72
С грамматикой			
bare	45.78	0.44	53.78
safe	93.21	0.74	6.058
randReading	95.50	2.59	1.90
freqReading	95.73	2.40	1.88
randReading+guessSyll	95.92	3.33	0.74
freqReading+guessSyll	96.15	3.14	0.72

Как видно нейросетевая модель успешно справляется с использование

контекста для расстановки ударений в омографах. Недостатками же представленной модели является очень простая архитектура и отсутствие работы с текстом. Далее мы будем использовать эту модель как базовую для сравнения результатов.

Глава 2.
Экспериментальная часть

2.1. Используемы данные

Во всех экспериментах в качестве данных мы использовали база данных акцентологической разметки в составе национального корпуса русского языка [15]. С каждым предложением в тексте были произведены следующие преобразование

1. Все буквы приведены к строчным.

2. Предложение разбито на смысловые подпредложение, используя в качестве разделителей знаки препинания. На этом этапе все знакия препинания удаляются.

3. Подпредложение содержавшие символы кроме кириллических букв удалены.

4. По правилам эксперимента из получившихся подпреложений собирались фразы.

Итоговые данные состояли из 3285455 слов. Все данные были разделены на 3 части: обучающая выборка (2299818 слов), валидационная выборка, применяемая для подбора параметров во время обучения модели (49281 слов) и тестовую выборку, на которые измерялся конечный результат (936356 слов). Омографы среди всех слов в нашей выборке составляют 3.31%. Распределение долей слов по длинам представлено в табл. 2.1. Аналогичное распределение для омографов представлено в табл. 2.2.

Таблица 2.1 — Распределение слов по числу слогов

Число слогов	Доля слов
2	0.474
3	0.308
4	0.144
5	0.053
6	0.015
7	0.003
8	0.001
9	10^{-4}

Таблица 2.2 — Распределение омографов по числу слогов

Число слогов	Доля слов
2	0.736
3	0.212
4	0.051

2.2. Метрики

Основной метрикой используемой для оценки окончательного качества является *Accuracy score* (2.1).

Обозначим позицию ударения во фразе как y_i. Позицию ударения, предсказанную моделью обозначим как y_i^*. Число фраз в выборке обозначим как N

$$ACC = \frac{\sum_{i=1}^{N} I\left\{y_i = y_i^*\right\}}{N} \tag{2.1}$$

2.3. Локальная модель

В качестве самой простой нейросетевой архетектуры нами была выбрана данная.

2.3.1. Архитектура модели

К входным данным применяется двустронняя рекурентная нейронная сеть на основе LSTM-модулей. Далее к каждому промежуточному вектору применяется один и тот же полносвязный слой с softmax активацией с размерностью выхода 2, первую компоненту этого вектора мы интерпретируем, как вероятность того что в данной позиции нет ударения, в вторую, как то что оно есть. Архитеткура представлена на рис. 2.1. Эту архитектуру далее мы будем называть локальной моделью.

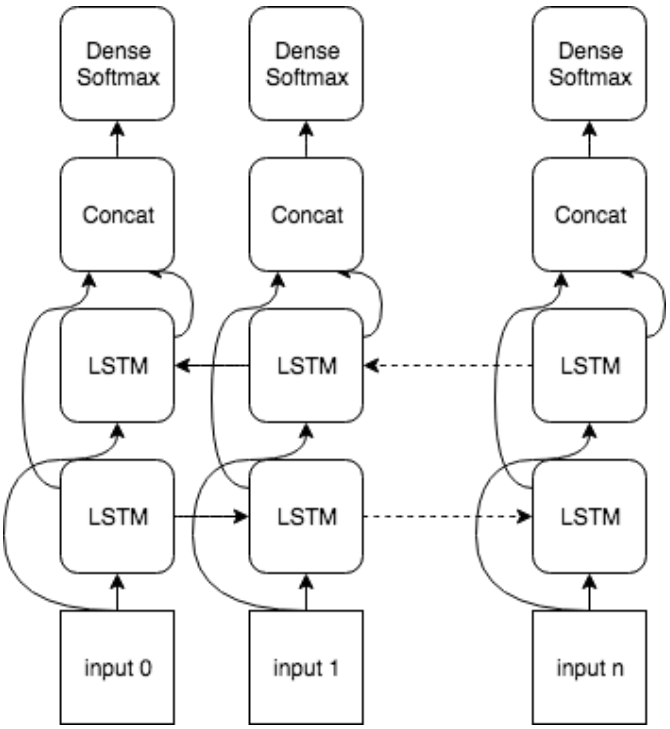

Рисунок 2.1 — Архитектура сети

2.3.2. Посимвольный эксперимент

Для получение более подробной картины того, как разные архитектуры нейронной сети и предобработка данных влияет на качество мы провели данный эксперимент, как упрощение базовой модели [14], которую мы будем называть глобальной моделью .

Входные данные: На вход модели подается слово, в котором мы хотим поставить ударение, 4 последние буквы предыдущего слова, если оно есть и 4 последние буквы следующего слова, если оно есть. 4 буквы мы используем, потому что это длина окончания в русском языке. Окончание

может нам помочь определить форму слова, что необходимо для удаления неоднозначности при расстановки ударения в большинстве омографов.

Позиция ударения во фразе выбиралась, как позиция с максимальной вероятностью второго класса.

Результаты этого эксперимента и сравнение с базовой моделью представлены в табл. 2.3

Таблица 2.3 — Сравнение результатов локальной и глобальной символьной модели

Число слогов	Локальная модель	Глобальная модель
Все слова		
2	0.961	0.983
3	0.940	0.977
4	0.947	0.976
5	0.960	0.977
6	0.958	0.973
7	0.924	0.955
8	0.866	0.923
9	0.809	0.952
среднее	0.952	0.979
Омографы		
2	0.839	0.810
3	0.774	0.844
4	0.787	0.847
среднее	0.821	0.819

Наше предположение о том, что эта модель более слабая, чем глобальная подтвердилось. При этом благодаря изменению использования контекста, результаты на омографах удалось улучшить

2.3.3. Эксперимент с предложением

Для исследования влияния длины контекста на качество расстановки ударений был проведен следующий эксперимент. Контекстом здесь является не окончания соседних слов, а все подпредложение (часть предложения между знаками препинаниия). Если введение в контекст других слов кроме соседних является тоже значимым, мы поймем это в этом эксперименте.

Входные данные: На вход модели подается подпредложение, описание построение находится в разд. 2.1. При этом модель должна расставить ударения во всех словах в подпредложении.

Для получения итогового результата из вектора с вероятностями мы в каждом слове выбирали символ с наибольшей вероятностью того, что на него падает ударение. Это в отличие отсечения по границе позволяет добиться того, что в каждом слове находится ровно одно ударение.

Результаты этого эксперимента и их сравнение с локальной символьной моделью представлены в табл. 2.4

Таблица 2.4 — Сравнение локальной символьной модели и локальной модели по предложениям

Число слогов	Символьная модель	Модель по предложениям
Обычные		
2	0.985	0.897
3	0.972	0.891
4	0.972	0.902
5	0.976	0.927
6	0.977	0.925
7	0.947	0.898
8	0.899	0.855
9	0.843	0.647
среднее	0.978	0.898
Омографы		
2	0.950	0.831
3	0.832	0.754
4	0.843	0.775
среднее	0.932	0.812

2.3.4. Слоговая модель

В русском языке ударение может падать только на гласные буквы. Модели приходилось также учитывать то, что на согласные буквы ударение падать не может. Из-за этого увеличивалась сложность модели, и количество информации, которое оно должна хранить. Также из-за этого увеличивалось время обучения.

Одним из вариантов решения этой проблемы является замена символьной модели на слоговую, то есть на вход модели будут подаваться закодированные слоги, а не символы.

Деление на слоги слов в русском языке однозначно установлено [16], поэтому в преобразование данных будет детерменированно.

После преобразования получилось 14083 слога.

Входные данные: Формат входных данных аналогичен формату, примененному в локальной символьной модели.

Результаты этого эксперимента, а также их сравнение с результатами локальной и глобальной сисвольной моделей представлены в табл. 2.5

Таблица 2.5 — Сравнение результатов локальной и глобальной символьной модлелей с локальной слоговой моделью

Число слогов	Слоговая модель	Локальная модель	Глобальная модель
Все слова			
2	0.985	0.961	0.983
3	0.972	0.940	0.977
4	0.972	0.947	0.976
5	0.976	0.960	0.977
6	0.977	0.958	0.973
7	0.947	0.924	0.955
8	0.899	0.866	0.923
9	0.843	0.809	0.952
среднее	0.978	0.952	0.979
Омографы			
2	0.950	0.839	0.810
3	0.832	0.774	0.844
4	0.843	0.787	0.847
среднее	0.877	0.821	0.819

Применение слогового кодирования позволило без изменения архитектуры модели повысить качество расстановки ударений на 2.6% для всех слов. Это позволило этой модели сравняться по качеству с глобальной символьной моделью. При этом качество расстоновки удаления в омографах у этой модели выше. Из этого всего можно сделать вывод, что обучение модели на слогах

вместо символов может помочь повысить результат без изменения архитектуры. Это можно объяснить тем, что модель учит векторные представления для слогов отдельно в embedding слое, а не пытается выучить подобные взаимосвязи в рекуррентном слое. Или в слоге содержится больше информации, чем в одной букве.

2.4. Глобальная модель

2.4.1. Архитектура модели

2.4.2. Слоговая модель

2.5. Attention

2.5.1. Архитектура модели

2.6. Conditional random field

2.6.1. Архитектура модели

2.7. Active learning

Заключение

Итоги работы

Дальнейшие исследования

Список использованных источников

[1] Еськова Н.А. *Словарь трудностей русского языка. Ударение. Грамматические формы.* М.: Языки славянской культуры, 2014.

[2] Kenneth Church. Stress assignment in letter to sound rules for speech synthesis. *Association for Computational Linguistics, 246–253*, 1985.

[3] Briony Williams. Word stress assignment in a text-to-speech synthesis system for british english. *Computer Speech and Language, 2:235–272*, 1987.

[4] Morris Halle. *On stress and accent in IndoEuropean.* Language, 1997.

[5] Keith Hall and Richard Sproat. Russian stress prediction using maximum entropy ranking. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 879–883, Seattle, Washington, USA*, 2013.

[6] Qing Dou, Shane Bergsma, Sittichai Jiampojamarn, and Grzegorz Kondrak. A ranking approach to stress prediction for letter-to-phoneme conversion. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 118–126, Suntec, Singapore*, 2009.

[7] Michael Collins and Terry Koo. Discriminative reranking for natural language parsing. *Computational Linguistics, 31:25–69*, 2005.

[8] Зализняк А. А. *Грамматический словарь русского языка.* М.: Русский язык, 1977.

[9] Robert Reynolds and Francis Tyers. Automatic word stress annotation of russian unrestricted text. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). Linkoping University Electronic Press, Sweden, Vilnius, Lithuania*, 2015.

[10] Kimmo Koskenniemi. Two-level morphology: A general computational model for word-form recognition and production. *Technical report, University of Helsinki, Department of General Linguistics*, 1983.

[11] Kenneth R. Beesley and Lauri Karttunen. *Finite State Morphology: Xerox tools and techniques.* CSLI Publications, Stanford, 2003.

[12] Fred Karlsson. Constraint grammar as a framework for parsing running text. *Proceedings of the 13th Conference on Computational Linguistics (COLING), 3:168–173, Helsinki, Finland*, 1983.

[13] Yulia Lavitskaya and Baris Kabak. *Phonological default in the lexical stress system of Russian: Evidence from noun declension.* Lingua, 2014.

[14] Maria Ponomareva, Kirill Milintsevich, Ekaterina Chernyak, and Anatoly Starostin. Automated word stress detection in russian. *Proceedings of the First Workshop on Subword and Character Level Models in NLP, 31–35, Copenhagen, Denmark*, 2017.

[15] Гришина Е. А. *Корпус «История русского ударения» // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы.* СПб.: Нестор-История, 2009.

[16] Литневская Е. И. *Русский язык: краткий теоретический курс для школьников..* М.:МГУ, 2006.