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 ражирования

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 его начала и концы (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).

| 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 |

Таблица 1.1: Абстрактные фонетические классы.

В качетсве данных авторы использовали Грамматический словарь русского языка[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.

| 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 1.2: Результаты ранжирования.

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 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, так как неправильное ударение может быть хуже, чем его отсутствие для человека.

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 их было несколько, то в зависимости от эксперимента выбиралась дальнейшая стратегия.

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

Описание экспериментов:

- **bare:** при нескольких возможных прочтениях слова, ударение в слове не проставлялось.
- **safe:** при нескольких возможных прочтениях, ударение в слове выставлялось, если в них всех 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 для учитывания 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.

| 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 1.3: Результаты применения конечного преобразователя.

При использовании модели без формальной грамматики среди 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 угаданных слов. 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, при этом, так как это тексты для начинающих, их словарь скорее всего был достаточно мал. Также проведение экспериментов RandReading и FreqReading непонятно, 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 длина фразы, соответствующий распределению вероятностей поставить ударение в конкретной позиции.(представлена на рис. 1.1)

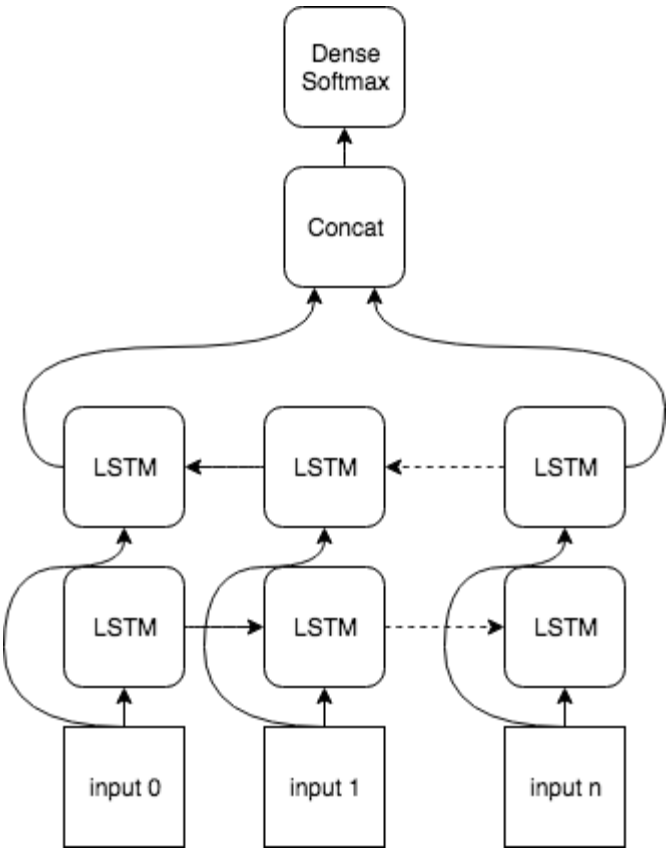


Рисунок 1.1: Архитектура сети

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.

| Эксперииент | 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 1.4: Результаты применения нейросетевой моделе 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 1.5: Результаты применения нейросетевой моделе 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 моделе на омографах.

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.
# Tmp

## 2.1. Основные обозначения и определения

Все говорят: Кремль, Кремль. Ото всех я слышал про него, а сам ни разу не видел. Сколько раз уже (тысячу раз), напившись или с похмелюги, проходил по Москве с севера на юг, с запада на восток, из конца в конец, насквозь и как попало – и ни разу не видел Кремля [**?**].

**Определение 1.** *Шизофазия (речевая разорванность) - симптом психических расстройств, выражающийся в нарушении структуры речи, при которой, в отличие от речевой бессвязности (потока несвязанных слов), фразы строятся правильно, однако не несут никакой смысловой нагрузки, а содержание речи соответствует содержанию бреда.[16]*

$$E = mc^2$$

**Обозначение 1.**
- $H$ — *водород.*
- $O$ — *кислород.*
- $C$ — *углерод.*
- ...

На формулки тоже можно ссылаться.

$$Women = Evil \tag{2.1}$$

Согласно (2.1), женщины — зло.

Собсно, на все леммы, теоремы, примеры, замечания и тэдэ можно ссылаться.

**Пример 1.** *Вот, например, Лёшка хотел Отл(10), а Иванова ему 9 поставила.*

**Замечание 1.** *Родился на улице Герцена, в гастрономе номер двадцать два. Известный экономист, по призванию своему — библиотекарь. В народе — колхозник. В магазине — продавец. В экономике, так сказать, необходим. Это, так сказать, система... э-э-э... в составе ста двадцати единиц. Фотографируете Мурманский полуостров и получаете «Те-ле-фун-кен». И бухгалтер работает по другой линии — по линии библиотекаря. Потому что не воздух будет, академик будет! Ну вот можно сфотографировать Мурманский полуостров. Можно стать воздушным асом. Можно стать воздушной планетой. И будешь уверен, что эту планету примут по учебнику. Значит, на пользу физике пойдёт одна планета. [16]*

В силу прим. 1 и замеч. 1, динозавры вымерли.

### 2.2. Ещё какая-то муть

Так. Стакан зубровки. А потом – на Каляевской – другой стакан, только уже не зубровки, а кориандровой. Один мой знакомый говорил, что кориандровая действует на человека антигуманно, то есть, укрепляя все члены, ослабляет душу. Со мной почему-то случилось наоборот, то есть душа в высшей степени окрепла, а члены ослабели, но я согласен, что и это антигуманно. Поэтому там же, на Каляевской, я добавил еще две кружки жигулевского пива и из горлышка альб-де-дессерт.[**?**]

# Глава 3.
# Заметки о женской логике

## 3.1. Как говорил Бек

В наш век точное познание завоевывает все новые области. Одна из таких областей – женская логика. Строгое изложение находится еще в стадии зарождения. Обычная мужская логика прошла эту стадию более двух тысяч лет назад, но женская логика еще ждет своего Аристотеля. Потомкам принадлежит большая и почетная задача создать систематический курс женской логики, выполнить ее аксиоматизацию, создать вычислительные машины, действующие по женским логическим схемам.[?]

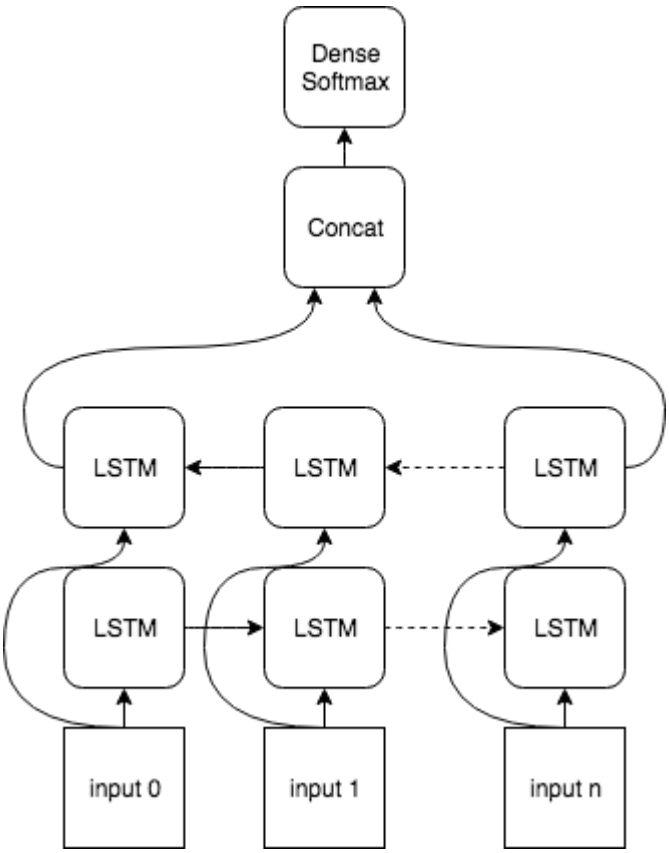


Рисунок 3.1: Чёрный квадрат

На картинку тоже можно ссылаться: рис. 3.1

## 3.2. Теорема Сосницкого

Следующая теорема даёт ответы на все ваши вопросы.

**Теорема 1** (Теорема Сосницкого)**.** *Lorem ipsum dolor sit amet, consectetur adipiscing elit.*

Для доказательства теоремы понадобится следующая лемма:

**Лемма 1.** *Если в кране нет воды, значит выпили жиды.*

*Доказательство.* Если в кране есть вода, значит жид нассал туда.

□

1. Пер.

2. Пер.

3. Пер.

**Следствие 1.** *О, тщета! О, эфемерность!*

*Доказательство.* О, самое бессильное и позорное время в жизни моего народа – время от рассвета до открытия магазинов! Сколько лишних седин оно вплело во всех нас, в бездомных и тоскующих шатенов! □

# Глава 4.
# Как рисовать всякие красивости

## 4.1. Система нумерованных уравнений

$$\frac{dq}{dt} = \frac{dH}{dp} \tag{4.1}$$

$$\frac{dp}{dt} = -\frac{dH}{dq} \tag{4.2}$$

А всё для того, чтобы сослаться раз (4.1), сослаться два (4.2)

| | | | |
|---|---|---|---|
| Р | О | К | К |
| Е | Б | О | Л |
| М | У | П | Ю |
| О | В | И | Ч |
| Н | И | R | Е |
| Т абсв | | | Й |

Таблица 4.1: Sample text.

Что бы вы думали можно сделать с разд. 4.1

## 4.2. Дерево

Хер знает зачем, но вдруг пригодится.

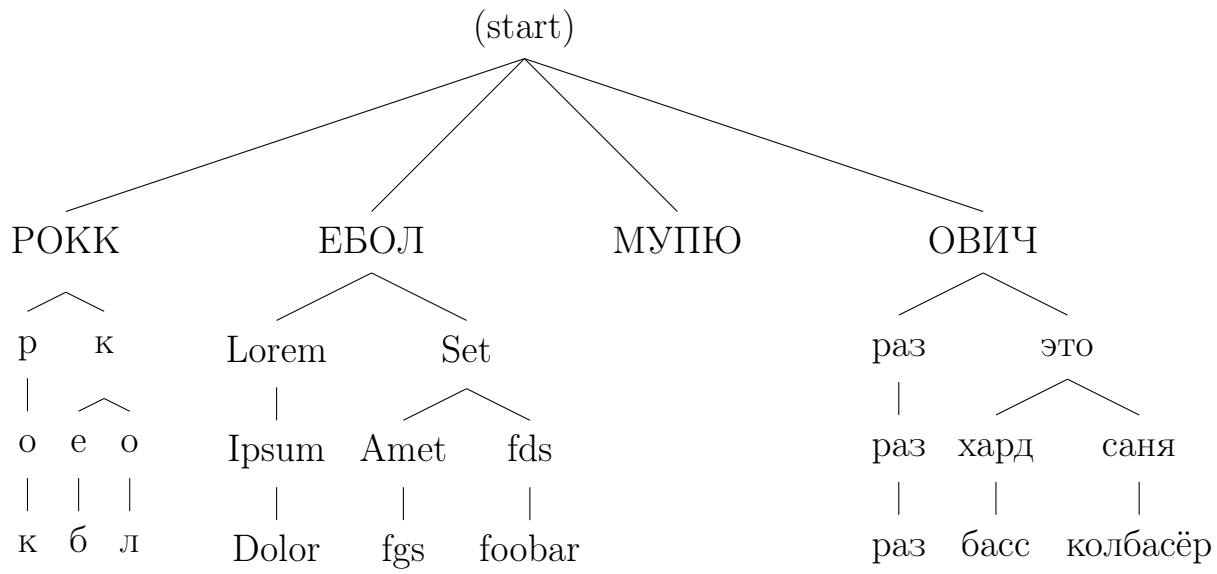


Рисунок 4.1: Смотри как умею

Опа рис. 4.1.

## Заключение

Несмотря на то, что можно это объяснить по-разному, хотя и в растерянности относительно возможных непониманий того, что следует из того, почему современные госпитали, и притом многие, с определенным и, вместе с тем неопределенным недоверием не выразили почти никакой заинтересованности в таком деле, как местная анестезия, и с полной уверенностью я при данных обстоятельствах не думаю, что стоит пытаться защищать не защищаемую репутацию хирургии вместо того, чтобы постараться привлечь на свою сторону других всяких разных, и это побудило меня несколько месяцев тому назад написать на эту тему большую часть чего-то вроде частично исчерпывающей статьи, закончить которую мне помешало плохое здоровье...

## 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] Неизв. Шизофазия. https://ru.wikipedia.org/wiki/Шизофазия, 1962.