Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2022»

Москва, 15–18 июня 2022 г.

Errors in the Russian Web Corpus

Maria Khokhlova
St. Petersburg State University,
St. Petersburg, Russia
m.khokhlova@spbu.ru

Abstract

The explosion of the Web leads to the production of large amounts of texts and inevitably influences their quality. Errors that tend to occur more often can distort results, especially when texts are used for scientific purposes, in language teaching or learning. Hence, there is a need to examine the existing corpora based on web texts and to clean up the data, which may contain such "noisy" fragments. In our study, we turn to the problem of errors and analyze the Aranea Russicum Maximum corpus. Among such errors, we can name, above all, encoding errors, incorrect font types, as well as segments written in other languages. These phenomena result in incorrect morphological analysis and lemmatization, frequency distortion, as well as the fact that lexical units cannot be found and therefore displayed to corpus users. The paper focuses on the errors, describe their types and outline possible ways to eliminate them.

Keywords: web corpora; Russian language; errors; typos; text quality
DOI: 10.28995/2075-7182-2022-21-1168-1176

О некоторых типах ошибок в русскоязычном Интернет-корпусе

Хохлова М. В.
Санкт-Петербургский
государственный университет,
Санкт-Петербург, Россия
m.khokhlova@spbu.ru

Аннотация

Лавинообразное развитие Интернета привело к тому, что качество веб-текстов значительно ухудшилось. Ошибки в них встречаются всё чаще, что не может не сказываться на результатах поиска, особенно, если эти тексты используются в научных целях: например, для создания корпусов или при обучении языкам. Таким образом, наблюдается необходимость изучить существующие корпусы, основанные на веб-текстах, и наметить способы улучшения их качества, если они содержат ошибки. В нашем исследовании мы обращаемся к проблеме ошибок и на материале корпуса Aranea Russicum Maximum. Среди таких ошибок можно назвать, прежде всего, ошибки кодирования, неверные символы, опечатки, а также фрагменты, написанные на других языках. Эти явления приводят к неправильному морфологическому анализу и лемматизации, частотным искажениям, а также к тому, что лексические единицы не могут быть найдены и, следовательно, отображены среди результатов. В работе дается анализ ошибок, описываются их виды и намечаются возможные пути их устранения.

Ключевые слова: Интернет-корпусы; русский язык; ошибки; опечатки; качество текстов

1 Введение

Лавинообразное развитие Интернета привело к тому, что качество веб-текстов значительно ухудшилось. Ошибки в них встречаются всё чаще, что не может не сказываться на результатах поиска, особенно, если эти тексты используются в научных целях: например, для создания корпусов или при обучении языкам. Таким образом, наблюдается необходимость изучить существующие

корпусы, основанные на веб-текстах, и наметить способы улучшения их качества, если они содержат ошибки.

Технологии в области создания корпусов существенно эволюционировали за последние двадцать лет. Объемы корпусов постоянно растут за счет обработки больших объемов текстовых данных, которые содержат преимущественно веб-тексты, поэтому важно обращать внимание на качество последних. Методы лемматизации, морфологического и синтаксического анализа изначально разрабатывались для стандартного (литературного) языка, поэтому их применение к иным текстам (в частности, полученным из Интернета) может приводить к ошибочным результатам. Ситуация становится хуже, если речь идет о «грязных» текстах, в которых содержится большое количество разного рода ошибок.

В особенности актуальным данный вопрос является в случае лингвистического анализа, т.к. при стандартном поиске в Интернете ошибки на веб-страницах могут не оказывать существенного влияния на результат, как это может быть в случае использования корпусов для лингвистических целей. Ошибки и опечатки могут привести даже к неожиданным результатам[1], так как их многократное повторение увеличивает вероятность этих явлений. Точность автоматических приложений, основанных на веб-корпусах, также может страдать из-за их возможного низкого качества. Кроме того, некоторые тексты получены с помощью оптического распознавания символов и не были вычитаны на следующем этапе, что приводит к наличию ошибок.

2 Обзор исследований

В течение последних нескольких лет Интернет-корпусы обрели большую популярность. Их объемы позволяют проводить разнообразные лингвистические исследования, а также обращаться к ним для проверки работы автоматических алгоритмов (например, исследование о влиянии размера корпусов на качество моделей эмбеддингов, построенных на их основе [Kutuzov, Kunilovskaya, 2018]).

Вопрос, связанный с отбором текстов и их обработкой, относится к методологии построения корпусов. В определенной степени он является решенным, когда речь идет о традиционных корпусах (об их отличиях от Интернет-корпусов см., например, [Хохлова 2016]). Однако в случае веб-текстов и построения коллекций большого объема данная проблема снова становится актуальной. Поэтапный процесс создания корпусов нового типа описан в работе [Jakubicek et al. 2020].

При разработке корпусов может возникать следующий вопрос: могут ли они содержать тексты, написанные на других языках? Если да, то какой максимальной длины могут быть подобные фрагменты? По всей вероятности, для веб-текстов эти вопросы не ставились. Многочисленное дублирование содержания веб-страниц составляет определенную проблему для корпусов текстов и также заслуживает внимания [Benko 2019].

Проблема выявления опечаток поднималась в ряде работ разных авторов. Прежде всего, речь идет о системах автоматической проверки орфографии [Shavrina 2017]. Авторы [Shavrina, Sorokin 2015] используют вероятностную модель, основанную на расстоянии Левенштейна для исправления ошибок в текстах социальных медиа. К вопросу нормализации этого же вида текстов обращаются в работе [Clark, Araki, 2011]. В 2016 г. в рамках конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог» прошло первое в России соревнование по автоматическому исправлению опечаток SpellRuEval. Распределение орфографических ошибок в английском и немецком языках рассмотрено на материале веб-текстов в [Ringlstetter et al. 2006]. Опечатки влияют на качество Интернет-текстов, для оценки которого в [Baeza-Yates, Rello 2012] предлагается соответствующая мера. История опечаток в русском языке и многочисленные примеры приводятся в работе [Шерих 2004]. В работе [Шаповал 2009] автор рассматривает грамотность школьников и проводит различие между ошибками, которые возникают при написании от руки, и опечатками, появление которых связано с использованием технических средств[2].

[1] Широко известен следующий пример из русской литературы: в поэме С. Есенина «Чёрный человек» долгое время были принято ошибочное написание *на шее ноги* вместо *на шее ночи*, которое возникло из-за неверной трактовки буквы *ч* (при написании от руки *г* и *ч* выглядит похожим образом).

[2] Хотя опечатки можно найти и в более ранних текстах, тем не менее, большое распространение они получили в связи с использованием компьютеров и множественным копированием одних и тех же фрагментов.

3 Материал исследования

В нашем исследовании мы обратимся к анализу русскоязычного корпуса Araneum Russicum III Maximum (19 778 млн словоупотреблений, или 1 078 млн слов) [Benko 2014]. Для нахождения типичных ошибок мы рассмотрим идентификатор языков [Гиляревский, Гривнин 1965], в котором описаны разнообразные алфавиты. В рамках нашего исследования мы сравним те символы кириллицы, которые совпадают с буквами других алфавитов. Также мы использовали список ошибок, приведенный в (Shavrina, Sorokin, 2015).

4 Типология ошибок

Корпусы текстов содержат ошибки разных типов, которые могут быть разными по своей структуре и по причине возникновения. Например, ошибки могут появляться вследствие неточного автоматического распознавания, неверных лемматизации или морфологического анализа. Тексты, которые загружаются в корпус, могут быть изначально низкого качества. Ниже мы приводим некоторые типы ошибок, которые могут быть найдены в корпусах:

1) орфографические ошибки (могут быть вызваны недостаточной языковой компетенцией пишущего или его намерением);
2) типографические (графематические) ошибки;
3) ошибки кодировки;
4) ошибки неверной процедуры автоматического распознавания текстов (OCR).

В своем исследовании мы обратимся к трем группам: графематическим ошибкам и ошибкам кодировки, а также к тем ошибкам, которые вызваны возможным неверным распознаванием отсканированных текстов. Опечатки связаны с порождением текста и встречаются, например, на форумах или в чатах, когда пишущий пытается быстро отреагировать, и допускает ошибки. Третья группа ошибок затрагивает художественную или научную литературу, потому что, как правило, эти тексты не создаются онлайн, а загружаются после процедуры распознавания. В этом случае ошибки вызваны схожестью между буквами. Например, русские буквы *п* и *н*. Мы оставляем за скобками вопрос об орфографических ошибках, которые связаны с уровнем грамотности. Мы также не будем рассматривать намеренные искажения, когда ошибки в правописании допускаются намеренно с некоторой целью.

4.1 Комбинации с расширенной кириллицей

Наиболее многочисленный тип ошибок, который мы смогли определить на данный момент, относится к тем случаям, когда происходит смешение букв расширенного кириллического и русского алфавитов. В этих случаях мы можем говорить или об иностранных словах, или об ошибках в русских словах, которые возникают из-за неверной процедуры распознавания символов. Подобные дополнительные символы используются в абхазском, азербайджанском, белорусском, болгарском, украинском и других языках. В довольно большом числе примеров эти буквы перепутаны с похожими русскими. В запросах к корпусу были использованы буквы иных алфавитов, которые схожи по своей графике с русской кириллицей. Результаты представлены частично в Табл. 1. Например, буква *р* используется вместо буквы *р* или буква ҷ заменена на букву *ч*.

Неверная буква	**Правильная буква**	**Абсолютная частота**	**Относительная частота в ipm**	**Примеры с неверным написанием**	**Примеры с правильным написанием**
ґ	г	6 725	0,34	ґраницы, временныґе, виґдение	границы, временные, видение
ҽ	е	49	0,00	жизнедеятҽльности, экстрҽнной, внутрҽнних	жизнедеятельности, экстренной, внутренних
ҏ	р	704	0,04	приватизации, адресата, словаре	приватизации, адресата, словаре
ҭ	т	1 496	0,08	размышляют, незаметно, культуры	размышляют, незаметно, культуры
ҷ	ч	1 089	0,06	аллегорических, ҷтобы, мальчик	аллегорических, чтобы, мальчик
ԁ	л	19	0,00	ԁаԁҽко, озԋа́чаԁо, ԋа́чаԁось	далеко, означало, началось
ԋ	н	1275	0,07	ԋе, сԋа́чала	не, сначала
е	е	11 042	0,60	ничего, несчастных, следующее	ничего, несчастных, следующее
г	г	1 814	0,09	критического, детского, гармонического	критического, детского, гармонического
ҫ	с	90	0,00	повторноҫтью, доҫтаточная, инҫтанции	повторностью, достаточная, инстанции
Є	Э, е, ё	11 995	0,61	Єто Єлизавета тЄмные	Это Елизавета тёмные (темные)
S	Б	4 803	0,24	Sолее, Sыл Sыстро,	Более Был, Быстро
s	разные	8 483	0,43	культурыs spri sочему	культуры при почему
J	ё	10 755	0,54	*Jлочка серьJзным* тJмными	*Ёлочка* серьёзным тёмными
љ	none, ль	3 270	0,17	неделиљ жизнедеятҽльности самостоятҽльно	недели жизнедеятельности самостоятельно
Ђ	none, dash	19 968	1,01	знакомство вЂ " отношение между людьми	знакомство — отношение между людьми

Таблица 1: Неверное использование символов в русских словах

Буквы других алфавитов могут обозначать иностранные слова и, таким образом, определение подобных случаев написания может свидетельствовать о том, что в корпусе содержатся большие фрагменты нерусскоязычных текстов. Например, символ *ґ* может либо неправильно использоваться в русских словах, либо идентифицировать лексемы, написанные на украинском языке

(например, *ґрунтуватися*, *ґатунку*, *ґвалтують*). Отдельные буквы могут указывать на бессмысленные и «грязные» тексты. Например, буква *È* встретилась в 191 слове, большая часть из которых содержит и другие ошибки, связанные с неверным распознаванием, кодировкой или изначальным низким качеством отрывков. Вышеупомянутые ошибки различаются по своей частоте, некоторые из них широко распространены в текстах, например, отрицание не неправильно записано как *иԑ* имеет 1255 вхождений (около 98% всех случаев, встречающихся с символом *и*). Опечатки в корпусе были неправильно лемматизированы, что привело к частотным искажениям.

4.2 Комбинирование латинских и кириллических символов

Схожий тип ошибок может быть найден в словах, которые содержат символы из двух алфавитов: кириллические символы заменяются на идентичные по своему написанию латинские и наоборот. В отличие от предыдущего случая, написание букв полностью совпадает и его невозможно отличить от верного. Такой подход иногда используется пользователями, чтобы затруднить индексацию поисковыми системами (например, при написании фамилий или адресов). Следующие буквы кириллицы имеют латинские строчные и прописные аналоги: *а*, *е*, *о*, *р*, *с*, *у* (прописная буква в этих двух алфавитах различна) и *х*. Заглавные буквы одинаковы для следующих букв: *В*, *К*, *Н* и *Т*. Например, *скромный*. Его написание на кириллице графически совпадает с вариантом *скромный*, встречающимся в корпусе, где используется латинская буква с. Несмотря на то, что оно правильно аннотировано как прилагательное, все 9 вхождений этого лексического элемента содержали одну и ту же лемму с ошибками. Системы Яндекс и Google исправляют опечатки в запросах и предупреждают о них, если пользователь скопирует подобные слова из результатов выдачи корпуса и ищет их при помощи поисковых систем. Однако такая процедура не всегда приводит к желаемому результату. Например, аббревиатура *ДНК* с опечаткой в виде латинской *Н* вместо кириллической *Н* будет преобразована в *ДРК*. Системы автоматически заменяют неверные буквы на те, которые расположены на той же клавише (в данном случае, на русскую *Р*). Следовательно, результаты будут показаны для ДРК, что означает «*Демократическая Республика Конго*». По предварительной оценке, в корпусе содержится около 2 млн примеров, написанных как кириллическими буквами, так и хотя бы одним из их вышеперечисленных латинских аналогов, например, *серебро*, *роса*, *свергнуть*, *ожидать*, *верхом*, *Некоторые* и т.д. Их можно обнаружить с помощью регулярных выражений как с русскими символами, так и с определенным набором латинских. Например, *[а-я]+[аоерсух]+* найдет 311 263 слова, оканчивающихся на нерусские символы, многие из которых имеют неправильные леммы и морфологические теги. Однако примеры этой многочисленной группы требуют дальнейшего анализа.

4.3 Комбинирование строчных и прописных букв

Комбинирование строчных и прописных букв также составляет проблему для последующих процедур лемматизации и морфологической разметки. Однако этот тип ошибок требует более пристального внимания. В ряде случаев достаточно конвертировать написание к одному регистру (обычно к нижнему), чтобы слова были корректно распознаны и лемматизированы. В других случаях верхний регистр может указывать на ударение (например, «бОльший»), особенности произнесения (например, «сердеШный») или аббревиатуры (например, «мАч» или кГц).

Ошибки могут встречаться на границе слов в тех случаях, когда намеренно или по ошибке не используются пробелы: *посмотретьДепортация* вместо «*посмотреть. Депортация*».

4.4 Неверная раскладка клавиатуры

Неверная раскладка, которая используется при наборе текста, приводит к тому, например, что русскоязычные слова написаны при помощи латинских символов. Несмотря на то, что автоматические системы переключают раскладку с кириллической на латинскую и наоборот, такие ошибки встречаются в текстах чатов или комментариев, а также в начале текстов. Так, *ghbdtn* используется вместо «привет» (65 случаев), *rfr* — вместо «как» или *ltkf* вместо — «дела» (часть конструкции «Как дела?», 120 случаев). Указанные примеры аннотированы в корпусе при помощи тега Z, который используется для обозначения пунктуации. Интересен факт, что подобные примеры встречаются в текстах, посвященных системам key switcher — автоматическим переключателям раскладок клавиатуры.

Если мы используем этот тег с регулярными выражениями для поиска лемм, написанных при помощи латинских символов, можно, таким образом, найти «грязные» тексты. Мы смогли найти более чем 153 млн примеров, которые соответствуют образцу *[atag="Z.*"&lemma="[A-Za-z]*"]*. Это слова, написанные латиницей, которые размечены как знак препинания. Они соответствуют иноязычным словам, например, *Yahoo*, *Corporation*, *Michael*, *email*, но также словам с опечатками, как в указанных примерах.

Данный тип возможных ошибок сложно поддается выявлению, т.к. при его поиске среди результатов продемонстрированы слова на иностранных языках. В качестве возможного варианта нахождения «подозрительных» случаев можно предложить искать комбинации букв, которые не встречаются в определенном языке, и тем самым отфильтровать нерелевантные тексты.

4.5 Использование дефисов

Некорректное использование дефисов может быть связано с неправильным распознаванием символов, когда слова переносятся через дефис в конце строк. Например, *загадоч-ный* вместо «*загадочный*». В корпусе 6 986 примеров (0,35 ipm) такой ошибки (слов с дефисным окончанием *-ный*[3]). Несмотря на то, что они были неправильно лемматизированы с дефисами, такие слова имели правильные морфологические теги A, соответствующие прилагательным. Тем не менее, неправильное использование дефисов или тире может вызвать не только ошибки лемматизации или морфологического анализа, но и неправильную токенизацию. По нашему мнению, данные ошибки могут быть откорректированы в определенной степени. Правила русской орфографии и пунктуации [1956] предполагают закрытый список элементов, которые следует писать через дефис (типа *-таки*, *-ка*, *-нибудь* и т.д.), и поэтому те случаи, которых нет в списке, могут расцениваться как возможные ошибки, возникающие на разрывах строк, и могут быть исправлены. Хотя необходимо понимать, что данный перечень не будет исчерпывающим и не будет включать случаи, относящиеся к авторскому использованию дефиса, или неологизмы.

4.6 Специальные символы

Ещё одна сложность, с которой сталкивается пользователь при работе с «грязными» текстами, заключается в использовании специальных символов. Например, символ ¶ показывает 11 212 вхождений (0,60 ipm) в корпусе. Фрагменты кодировки HTML также имеют множество примеров, среди которых можно найти неразрывный пробел (& nbsp) или разные типы дефисов (&ndash или &dash). Таким образом, на   приходится 195 266 обращений (9,9 ipm), а на &ndash/&mdash — 7 044 (0,36 ipm).

Для исправления подобных ошибок можно составить список символов или их комбинаций, которые необходимо отфильтровать. Специальные символы, которые отображаются среди результатов, могут сочетаться с другими ошибками и таким образом могут помочь выявить нерелевантные тексты. Например, они могут указывать на неправильную кодировку документа.

4.7 Опечатки

Данный подтип ошибок связан с тем, что буквы на клавиатуре расположены рядом, поэтому пользователь по ошибке нажимает на неверную клавишу или на две клавиши одновременно[4]. Например, предлог «для» может быть напечатан как *ддя*, поскольку буквы «д» и «л» расположены на стандартной русскоязычной раскладке рядом. В корпусе насчитывается 619 подобных примеров (0,03 ipm), которые получили неверную частеречную разметку существительного или глагола вместо предлога.

В некоторых случаях речь идет о лишних буквах, которые появляются в словах, или о пропуске. Например, форма *гооворить* встретилась 18 раз, в то время как *сказвать* имеет 4 вхождения в корпусе. Подобные ошибки могут касаться имен собственных: например, *Петрогад* вместо «Петроград», *Петербуг* вместо «Петербург».

---

[3] Нахождение всех подобных случаев связано с определенными сложностями. Решение могло бы заключаться в описании перечня морфем и их возможных комбинаций, которые переносятся на новую строку, при помощи регулярных выражений.

[4] Хотя также можно предположить, что подобные ошибки являются следствием неверного распознавания текстов.

Также может встретиться перестановка букв, когда из-за большой скорости печати пишущий нажимает клавиши в ином порядке следования: например, *гооврить* (75 случаев), предлог *вмсето* (54 случая), *проглука* (8 случаев), *дург* (138 случаев).

Опечатки сложно найти, т.к. они могут встретиться в любом слове, хотя можно отметить некоторую регулярность в определенных случаях. Так, буква *ю* имеет тенденцию ошибочно использоваться в конце предложений вместо обычных символов пунктуации (точки или запятой). Например, *сказатью* вместо «сказать» (6 случаев). В то время как *z* появляется в конце слов на месте *я* (например, *содержанiz*).

Возможным решением может быть использование программ для проверки правописания, чтобы откорректировать слова, которые не найдены в словарях.

5 Обсуждение результатов

Шаврина и Сорокин [Shavrina, Sorokin 2015] используют расстояние Левенштейна только для слов, которые отличаются не более чем двумя символами. Тем не менее нам удалось найти примеры с более чем двумя опечатками. Например, существительное «жизнедеятельность» имеет пять опечаток в варианте *жизнедеятѣльность*. Как справедливо отмечают авторы, расстояние Левенштейна может помочь избавиться от некоторых опечаток.

На основании рассмотренных и проанализированных нами примеров можно сделать вывод, что в большинстве своем те тексты, которые содержат вышеописанные ошибки, характеризуются низким качеством и выявляют большое количество других ошибок. Можно предположить, что частота этих единиц для такого большого русского корпуса незначительна; однако при рассмотрении языковых явлений мы часто имеем дело с малыми частотами, поэтому важно получать и изучать «чистые» данные.

Анализ показывает, что в ряде случаев буквы в словах с опечатками могут быть заменены обычными (традиционными), присущими русскому языку. Помимо слов, в которых ошибочно смешиваются символы из каких-то двух алфавитов (латинского и русского, расширенной кириллицы и русского), встречаются ошибки кодировки, иностранные слова, бессмысленные последовательности и т. д. Кроме того, отметим, что тексты, написанные на разных языках, оказываются разной длины. Например, фрагменты на удмуртском языке короткие, тогда как украинские или сербские слова встречаются в более крупных текстах.

Процедуры, связанные с очисткой текстов, при необходимости могут включать этапы, связанные с фильтрацией ошибок. Возможным решением может быть создание предварительного перечня веб-страниц, тексты которых могут войти в корпус. В таком случае корпус будет состоять только из проверенных текстов.

6 Заключение

Мы рассмотрели наиболее распространенные ошибки и обозначили несколько способов их решения. «Зашумленные» тексты можно очистить, удалив опечатки, которые встречаются регулярно, или переписав их кириллицей. Предлагаемый список ни в коем случае не является полным и исчерпывающим. Описанные примеры представляются нам достаточно интересными, хотя и не столь многочисленными. По предварительным оценкам, они составляют не более 1,5% от общего объема корпуса Aranea Russicum Maximum. Тем не менее, считаем важным уделить внимание вопросу очистки данных в рассмотренном корпусе, а также удаления из него нерелевантной информации. Полученные нами результаты могут быть использованы для построения новых корпусов.

Также в будущем было бы интересно посмотреть, существует ли корреляция между ошибками определенного типа и типом текста, его жанром, тематикой веб-сайтов или иными характеристиками. Мы также считаем, что наша работа позволит устранить уже найденные ошибки и, следовательно, подготовить более качественные корпусы.

Благодарности

Исследование выполнено при финансовой поддержке Санкт-Петербургского государственного университета (проект № 92563238).

References

[1] Baeza-Yates R., Rello L. (2012). On measuring the lexical quality of the web. In Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality. Lyon, France. Available at: https://dl.acm.org/doi/pdf/10.1145/2184305.2184307

[2] Benko V. (2014). Aranea: Yet another family of (comparable) web corpora. In International Conference on Text, Speech, and Dialogue, Springer, 247–256.

[3] Benko V. (2019) Deduplication in Large Web Corpora. In Bański, Piotr/Barbaresi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Kupietz, Marc/Lüngen, Harald/Iliadi, Cadoline (Eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22 July 2019. – Mannheim: Leibniz-Institut für Deutsche Sprache, 2019. Pp. 17-21.

[4] Clark E., Araki K. (2011). Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. In Procedia — Social and Behavioral Sciences, 27, 2–11.

[5] Gilyarevskiy R.S., Grivnin V.S. (1965). Language identification guide based on their systems of writing [Opredelitel' yazykov mira po pis'mennostyam]. Moscow.

[6] Jakubicek M., Kovar V., Rychly P., Suchomel V. (2020). Current Challenges in Web Corpus Building. In Proceedings of the 12th Web as Corpus Workshop. Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020, 1–4.

[7] Khokhlova M.V. (2016). A survey of large Russian web corpora [Obzor bol'shikh russkoyazychnykh korpusov textov]. In Computational Linguistics and Ontologies: Proceedings of the 19th International Scientific Conference "Internet and Modern Society" [Komp'yuternaya lingvistika I vychislitel'nye tekhnologii: sbornik nauchnykh trudov. Trudy 19 mezhdunarodnoy ob'yedinennoy nauchnoy konferentsii "Internet I sovremennoye obschestvo" (IMS-2016)]. St. Petersburg: ITMO University, pp. 74–77.

[8] Kutuzov A., Kunilovskaya M. (2018). Size vs. structure in training corpora for word embedding models: Araneum Russicum maximum and Russian national corpus. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10716 LNCS. https://doi.org/10.1007/978-3-319-73013-4_5

[9] Rules of Russian Spelling and Punctuation [Pravila russkoy orfografii i punktuatsii]. Moscow: Uchpedgiz, 1956.

[10] Ringlstetter Ch., Schulz K., Mihov S. (2006). Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. In Computational Linguistics 32 (3), 295–340.

[11] Shapoval V.V. (2009). The new types of mistakes in written speech [Novye tipy oshibok v pismennoy rechi]. In Russkiy yazyk v shkole, 9, 76–83.

[12] Shavrina T.O., Sorokin A.A. (2015). Modeling advanced lemmatization for Russian language using TnT-Russian morphological parser [Modelirovaniye rasshirennoy lemmatizatsii dlya russkogo yazyka na osnove morfologicheskogo parsera TnT-Russian]. In Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog». Selegey V. (ed.). Moscow: Russian State Univ. for the Humanities, 2015. Available at: http://www.dialog-21.ru/digests/dialog2015/materials/pdf/ShavrinaTOSorokinAA.pdf.

[13] Shavrina T.O. (2017). Methods of misspelling detection and correction: a historical overview [Metody obnaruzheniya i ispravleniya opechatok: istoricheskiy obzor]. In Voprosy yazykoznaniya, 4, 115–134.

[14] Sherikh D.Yu. (2004). "A" has fallen, "B" has disappeared… A curious history of typos [«A» upalo, «B» propalo... Zanimatel'naya istoriya opechatok]. Access mode: http://www.speakrus.ru/mix/opechatki/

References

[1] Baeza-Yates, R. & Rello, L. (2012). On measuring the lexical quality of the web. In Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality. Lyon, France. Available at: https://dl.acm.org/doi/pdf/10.1145/2184305.2184307

[2] Benko, V. (2014). Aranea: Yet another family of (comparable) web corpora. In International Conference on Text, Speech, and Dialogue, Springer, 247–256.

[3] Benko, V. (2019) Deduplication in Large Web Corpora. In Bański, Piotr/Barbaresi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Kupietz, Marc/Lüngen, Harald/Iliadi, Cadoline (Eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22 July 2019. – Mannheim: Leibniz-Institut für Deutsche Sprache, 2019. Pp. 17-21.

[4] Clark, E. & Araki, K. (2011). Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. In Procedia — Social and Behavioral Sciences, 27, 2–11.

[5] Jakubicek, M., Kovar, V., Rychly, P. & Suchomel, V. (2020). Current Challenges in Web Corpus Building. In Proceedings of the 12th Web as Corpus Workshop. Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020, 1–4.

[6] Kutuzov, A., & Kunilovskaya, M. (2018). Size vs. structure in training corpora for word embedding models: Araneum Russicum maximum and Russian national corpus. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10716 LNCS. https://doi.org/10.1007/978-3-319-73013-4_5
[7] Ringlstetter, Ch., Schulz, K. & Mihov, S. (2006). Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. In Computational Linguistics 32 (3), 295–340.
[8] Shavrina, T.O. & Sorokin, A.A. (2015). Modeling advanced lemmatization for Russian language using TnT-Russian morphological parser [Modelirovaniye rasshirennoy lemmatizatsii dlya russkogo yazyka na osnove morfologicheskogo parsera TnT-Russian]. In Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog». Selegey V. (ed.). Moscow: Russian State Univ. for the Humanities, 2015. Available at: http://www.dialog-21.ru/digests/dialog2015/materials/pdf/ShavrinaTOSorokinAA.pdf.
[9] Гиляревский Р.С., Гривнин В.С. Определитель языков мира по письменностям. М., 1965.
[10] Правила русской орфографии и пунктуации : Утв. Акад. наук СССР, М-вом высш. образования СССР и М-вом просвещения РСФСР. Москва : Учпедгиз, 1956.
[11] Хохлова М.В. Обзор больших русскоязычных корпусов текстов // Компьютерная лингвистика и вычислительные онтологии: сборник научных статей. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2016), Санкт-Петербург, 22 – 24 июня 2016 г. СПб: Университет ИТМО, 2016.С. 74–77.
[12] Шаврина Т.О. Методы обнаружения и исправления опечаток: исторический обзор // Вопросы языкознания, № 4, 2017. С. 115–134.
[13] Шаповал В.В. Новые типы ошибок в письменной речи // Русский язык в школе, № 9, 2009. С. 76–83.
[14] Шерих Д.Ю. «А» упало, «Б» пропало... Занимательная история опечаток. Access mode: http://www.speakrus.ru/mix/opechatki/