Лавошникова Э.К.

Московский государственный университет им. М. В. Ломоносова, Научно-исследовательский вычислительный центр, литературный редактор журнала «Вычислительные методы и программирование: Новые вычислительные технологии», el.lavoshnikova@yandex.ru

MICROSOFT WORD И ПРИЧИНЫ ПРОПУСКА ОШИБОК

КЛЮЧЕВЫЕ СЛОВА

Word'2013, компьютерная проверка правописания, системный словарь, русский язык, орфографические ошибки, синтаксические ошибки, опечатки, текстовый редактор, спеллер, автокорректор, программа-подсказка.

АННОТАЦИЯ

Рассматривается проблематика компьютерных систем проверки правописания. Разбирается работа автокорректора (спеллера), встроенного в текстовый редактор Microsoft Word (версия 2013 г.). На многочисленных примерах показано, что перегруженность системных словарей устаревшей и низкочастотной лексикой приводит к пропуску ошибок. Предлагается дополнять систему списками словоформ с наиболее «популярными» ошибками и информацией об их правильном написании.

Введение

При компьютерной проверке правописания в текстовом редакторе MS Word (здесь мы рассматриваем тексты, написанные на русском языке) автокорректор красной волнистой линией подчеркивает слова (или даже части сложных слов при написании через дефис), отсутствующие или не порождаемые в Word'овских системных словарях. Тем самым пользователю предлагается обратить на них внимание – нет ли в них ошибки или опечатки. *Программа-подсказка* после замен, вставок, удаления и перестановок некоторых символов ищет в системных словарях полученные «похожие слова», т.е. цепочки букв. Если находит, то выдает по желанию пользователя список таких словоформ как возможные варианты исправления неопознанного слова.

Заметим, что спеллер ОРФО первых версий Word'овского текстового редактора базировался на 1-м издании «Грамматического словаря русского языка» Андрея Анатольевича Зализняка, размеченном при скудном, в отличие от типографского в словаре, наборе символов на ЭВМ ЕС-1022 и перенесенном на бобины с магнитной пленкой в 1980-х годах коллективом Лаборатории автоматизированных лексикографических систем Научно-исследовательского вычислительного центра МГУ им. М.В. Ломоносова [4, с. 32].

Словарь А.А. Зализняка впервые был издан в 1977 году и с тех пор неоднократно переиздавался. Электронная версия этого словаря легла в основу большинства современных компьютерных программ, работающих с русской морфологией. Следует отметить, что словарь Зализняка [3] входит в список четырех словарей, грамматик и справочников, рекомендованных в 2009 году Межведомственной комиссией по русскому языку при Минобрнауки и содержащих нормы современного русского литературного языка.

1.**Проверка синтаксиса и пунктуации в текстовом редакторе Word'2013**

В книге, содержащей ровно 256 страниц, Ирина Спира пишет: «Традиционная проверка правописания была реализована в Microsoft Word на высоком уровне. Программа замечала не только орфографические ошибки, но и «чувствовала» весьма тонкие грамматические и стилистические нюансы, решала даже непростые пунктуационные задачи. Но в Microsoft Word 2013 качество проверки правописания русского текста заметно ухудшилось» [10, с. 59].

Мы не будем в настоящей статье обсуждать качество проверки стилистики или пунктуации, которое вызывает много вопросов, только слегка коснемся проблемы отсутствия реакции автокорректора на неправильный синтаксис фраз. Но даже в наиболее разработанном направлении – выявлении орфографических ошибок – в Word'2013 еще остается немало разного рода недочетов (см. недавние публикации автора [5; 7]).

Возможно, разработчики Word'2013 отказались от многих рекомендаций своей программы-подсказки по стилистике, синтаксису и пунктуации из-за многочисленных огрехов, о которых мы (наряду с другими пользователями) неоднократно писали, *например в статье* [6].

Например: спеллер-2013 подчеркивает *синей* волнистой линией слово *например* в тех случаях, когда запятая после этого слова исказила бы смысл высказывания (см. первую фразу настоящего абзаца). Кроме того, не предусмотрена возможность двоеточия после слова *например* (как во второй фразе). А если слово «например» заключено в кавычки, то подсказка предлагает поставить запятую с пробелом после *открывающей* кавычки (т.е. **«, например»**)! Во всех этих случаях подсказка выдает сообщение: «Пропущена запятая после вводного слова или перед ним».

Кроме того, Word'2013 *всегда* требует запятую перед «как» – вопреки правилам русской грамматики.

2.Перегруженность системных словарей низкочастотной лексикой

Ниже приводятся примеры специально сконструированных фраз с достаточно вероятными ошибками и опечатками, пропускаемыми вордовским спеллером-2013. К синтаксису этих фраз никаких замечаний (подчеркиваний синей волнистой линией) тоже не выдается.

«Вы зеленя молод ежь, а мыслете стар чески». В результате пропуска буквы «а» получились *зеленя*; при разбиении слова «молодежь» возник императив («Ёжь!») от неупотребительного, но включенного в вордовский системный словарь глагола «ёжить»; «мысл**е**те» – «старое название буквы м»; из-за появления лишнего пробела образовалась форма существительного «чёска».

«Как только вы уед**и**те, всех трудящ**их** (!) пер**е** сел**е**м в общ**е ж**итие» (получились формы слов *уесть*, *трудить*, *перо*, *сель*, *обще*).

«Не хочу вдаваться в эти де**рб**и, но нам **е**кну, что з**ап**родаваемым **щ**елком стоят…» (здесь: «дебри» с перестановкой букв, формы глаголов «*ёкнуть*» и «запродавать», вместо существительного «шелк» – междометие «щелк»).

Случается, что рядом расположенные на клавиатуре «л» и «д», «м» и «и», а тем более слабо различимые «ш» и «щ» по невнимательности пользователя оказываются «взаимозаменяемыми». В нашей кириллице в разных шрифтах схожи между собой цифра 3, буквы «з» и «э», «ь» и «ъ» и некоторые другие.

Еще примеры *с заменой соседней по клавиатуре буквы* (и со вставкой или отсутствием пробела), не подчеркиваемые в редакторе Word 2013: «**д**опасти вентиля тора» («лопасти»), «с за**ж**ором **пре** возмогу…**»** («с задором», «пря», «возмочь»), «лыбиться» (просторечие, ср. «дыбиться»), «он сосе**ж**у — з**ад**руга и за брата» («я сосежу» – от малоупотребительного «соседить»), «один лист при пере **писи** населения уме**ш**ает…» («умещает»), «ухо**л**ить з**ав**одкой» (вместо «уходить за водкой»).

В системных словарях Word'овского автокорректора довольно много слов, которые пропускаются без подчеркивания, т.е. без указания на возможную опечатку, но *с большей вероятностью* могут появляться в текстах пользователя в результате *пропуска буквы* (например, при недостаточно сильном нажатии клавиши). Примеры: ***вз**мутиться, **вс**кользнуть, в**ыв**ить* (вероятнее пропуск буквы в слове *вы**я**вить*), *в**ыт**роить*, *замет**та*** (ср. *замет**к**а*), *з**ат**есняться, зацеп**а**, над**в**ить* (ср. *над**а**вить), н**ау**стить* (ср. *на**п**устить*), *отв**ее**т* (ср. *отве**д**ет*), *пов**ет**ь* (помещение в крестьянском дворе, обл.), *под**со**чить, **по**петый* (ср*. п**р**опетый*), *по**ду**стить* (ср. *под**п**устить*), *по**мс**титься, **сб**ирать*, ***ср**оить*, *уто**чи**ть.* И немало других подобных слов можно привести. О перегруженности Word'овских системных словарей низкочастотной и малоупотребительной в современных текстах лексикой, которая мешает компьютерному выявлению опечаток, мы писали в работах [5; 7; 8] и других.

➢Об именах собственных и их компьютерной проверке

Четвертое и последующие издания «Грамматического словаря» А.А. Зализняка с обратным алфавитным порядком (последнее слово «несовершеннолетнЯЯ») дополнены приложением «Имена собственные» (более 8 тыс. словарных статей) [3, с. 731]. Однако версии текстового редактора MS Word разных лет еще до выхода 4-го издания (т.е. до 2003 года) уже содержали в своих внутренних словарях некоторые имена, фамилии и топонимы.

Можно привести примеры, когда представительность системного словаря имен собственных мешает выявлять слова с ошибками. Допустим, в тексте есть фразы, начинающиеся словами «Да**на**я задача…» или «С**тр**ый вариант…» с нечаянным пропуском буквы. Но эти опечатки не будут замечены Вордом-2013, поскольку в его словари внесено имя *Даная* и топоним *Стрый* (река, город).

В некоторых из предыдущих версий Word'а гипокористические (уменьшительно-ласкательные) имена *Катя*, *Маша*, *Юля*, *Коля*, *Костя*, *Боря* пропускались в текстах без подчеркиваний только потому, что спеллер считал их деепричастиями от глаголов «катить», «махать», «юлить», «колоть», «костить» и малоупотребительного «бороть». Однако Word'2013 уже

включил эти уменьшительно-ласкательные формы в свой системный словарь имен собственных – это можно утверждать на том основании, что косвенные падежи (*Катям*, *Машей*, *Юлями, Костю*) пропускаются без замечаний, в то время как набранные строчными буквами «катям», «машей», «юлями», «костю» подчеркиваются красным. Однако без красного подчеркивания пропускаются «колей» (родительный падеж мн. ч. от «колея»), «коле» («о коле в дневнике»), «о боре» (предложный падеж слова «бор»), «борей» (северный ветер).

В Word'овском системном словаре имен собственных содержатся «неполные» имена Алик, Алька, Гарик, Гоша, Ивашко, Катюшка, Костик, Марусенька, Муся, Нюша, Петруха, Санька, Сашура, Стасик, Степашка, Танька, Тимошка, Шурочка, даже Чук и Гек (герои повести Аркадия Гайдара). Этот список далеко не полон.

«**Шу**рок развязался на ботинке у Гагарина (когда он шел рапортовать Хрущеву)». Вордовский спеллер-2013 эту фразу пропускает без замечаний – из-за включения в системный словарь имени *Шурка*. Еще пример: «Л**уш**е п**о е**хать Михайло вой, Петро вой и Данило вой». Здесь специально допущены опечатки: пропуск буквы «ч» и разбиение фамилий. Но у спеллера-2013 нет замечаний, так как в его системном словаре имеются имена *Луша*, *Михайло, Петро* и *Данило*. Кстати, *синтаксис* этого предложения (предлог «по» перед глаголом «ехать») у Ворда-2013 возражений тоже не вызывает.

Приведем примеры фраз с ошибочным появлением пробела внутри слова: «Люсь, перестань хны кать! Кать, я не упрям люсь». Из песни Высоцкого (но с пробелом в слове «магазин»): «…Всё время тянет в мага зин. Подвинься, Зин!» Спеллер подчеркивает красным только «кать», «люсь» и «зин» со строчной буквы. Звательные формы Люсь, Кать, Зин (примеры звательного падежа: «отче наш», «боже мой», «чего тебе надобно, старче») спеллер, очевидно, воспринимает как родительный падеж мн. числа, поскольку, например, звательные формы *Ваньк, Зойк,* в отличие от форм мн. числа *Ванек, Зоек,* подчеркиваются красным.

Еще пример: «Прош, Кеш, про**ш у**, не трогайте кеш». Во этой фразе вставлен пробел в слово «прошу», поэтому «прош» со строчной буквы подчеркивается красным – в отличие от «Прош» и «Кеш». Однако подчеркивается и форма «кеш» со строчной буквы – вопреки рекомендации академического словаря [9]. Разработчики спеллера правильным считают популярное (наверное, более частое) написание «кэш».

В вордовском текстовом редакторе содержится имя *Мишель*, но только с мужским вариантом склонения, хотя у нас всё чаще этим именем стали называть новорожденных девочек. Есть в системном словаре и непопулярное имя *Ульян*, а модного женского имени *Ульяна* нет – форма «Ульяной» подчеркивается красным.

Еще пример без подчеркиваний автокорректора: «Квашей хи баре нужно при смотреться». «Квашей» и даже «Кваш» пропускается (а «квашей» со строчной буквы подчеркивается красным). Может, актер Игорь Кваша имеется в виду? Текстовый редактор склоняет эту фамилию во множественном числе, а форму в ед. числе «Квашой» спеллер почему-то подчеркивает как неопознанную.

Автокорректор-2013 не замечает разбиений «**Влади** кавказский», «**Влади** слав» или «**Влади** лен» (немодное уже имя-сокращение от «Владимир Ленин») с ошибочной вставкой пробела – очевидно, по причине присутствия в системном словаре фамилии-псевдонима актрисы Марины Влади (и это *Влади* –не от уменьшительного имени Владя, подчеркиваемого красным).

Написание «Фили**п**» с одной буквой «п» пропускается спеллером без замечаний. Наверное, при включении в системный словарь имелся в виду киноактер Жерар Филип (1922–1959).

Если набрать «Козаки, казаки и козаки», то первое слово спеллером-2013 пропускается без замечаний, а последнее устаревшее написание со строчной буквы подчеркивается красным. То же самое произойдет с фразой «Братья Волчеки и их волч**е**к» (последнее слово подчеркивается спеллером, так как правильно – волч**о**к). Очевидно, что фамилии *Козак* и *Волчек* внесены в Word'овский словарь имен собственных. Фамилия Бочкар**е**в через букву «е» пропускается без замечаний, а «Бочкар**ё**в» почему-то подчеркивается красным.

Теперь приведем еще два примера специально придуманных фраз – с пропуском буквы «й»: «Придется лопат**о** копать». Здесь слово «лопато» подчеркивается красным (но первым в подсказке выдается вариант «**Л**опато»). Теперь переставим слова. Фразу «Лопат**о** придется копать» Word'2013 пропускает без замечаний. Причина – присутствие в системном словаре фамилии *Лопато* (смотрим в Википедии: «Георгий Павлович Лопато был главным конструктором первой ЭВМ, разработанной в СКБ завода им. Г.К. Орджоникидзе»).

Версии текстового редактора Word, как предыдущие, так и 2013 года, «не знают»

отыменного прилагательного *лермонтовский,* в то время как слова *пушкинский, гоголевский* и *толстовский* пропускаются без замечаний. В правильно написанном «Людвиг ван Бетховен» спеллер-2013 подчеркивает «ван» и при вызове подсказки предлагает среди прочего «Ван» с прописной буквы.

Не всем известно, что псевдоним американского писателя О. Генри пишется через точку, а не через апостроф [3, с. 757]. Однако в системных словарях Word'овского автокорректора имеется словарная единица «О'Генри».

В нашей действительности фамилии с разнообразными ошибками исправлению не подлежат, потому что «так написано в паспорте». Вот пример реальной фамилии – *Щастливый* (такое написание прилагательного встречается в старинных текстах, но оно противоречит *современным* правилам). Подсказка-2013 выдает единственный вариант «исправления» этой фамилии: «Растлевай»!

Из словаря Зализняка [3, с. 767] мы узнаем, что творительный падеж *русской* фамилии *Чаплин* (в этимологических словарях дается ее происхождение от слова *цапля*) – «Чаплин**ы**м» (но спеллер-2013 этот вариант подчеркивает красным), однако следует писать «с Чарли Чаплин**о**м» (поскольку он иностранец), а также что в творительном падеже должно быть «Игорем Северянин**о**м», а не «Северянин**ы**м». Неправильный вариант с прописной (заглавной) буквы и окончанием «ым» спеллером-2013 тоже пропускается – вопреки правилу склонения существительных *северянин*, *крестьянин, южанин* и т.п.

Еще примеры особенностей в склонении имен собственных: «с князем МышкинЫм», но «знакомимся с городом МышкинОм» («Мышкин**ы**м» пропускается без возражений, а форма «Мышкин**о**м» подчеркивается Вордом как неопознанное слово), «с Пушкин**ы**м на дружеской ноге», но «под городом Пушкин**о**м» (в этом случае – без подчеркиваний спеллера). В настоящее время есть город Пушкин (до 1918 г. Царское Село, с 1918 по 1937 – Детское Село) и город Пушк**ино** Московской области.

В словаре Зализняка [3, с. 739] по поводу названий населенных пунктов на «ово», «ево» и «ино» читаем: «…очень часто встречается – как в устной речи, так и в печати – употребление данного слова как неизменяемого <…> Степень распространения этого явления так значительна, что, по-видимому, оно уже приближается к статусу допустимого варианта». То же самое можно сказать о фамилиях типа *Шевченко* (см. [3, с. 739]), которые всё чаще уже не склоняются.

Владимир Андреевич Успенский в статье «Субъективные заметки о неправильной норме» [11, с. 539] пишет: «Автор [Успенский. – *Э.Л.*] склоняется к тому, что понятие **нормы** имеет в своей основе статистику: если «так говорит» или «так понимает» абсолютное большинство носителей языка, то это и есть **норма**. С противопоставлением '**правильно–неправильно**' дело обстоит сложнее: единого ответа на вопрос «Что есть истина?» дать, по всей видимости, невозможно. Тем не менее, каждый отдельный пример устного или письменного словоупотребления допускает оценку (скорее всего, субъективную) по шкале **'правильно–неправильно'**. Тогда, хотя бы теоретически, возникает возможность **неправильной нормы**, когда нечто неправильно, но все так говорят». (См. также статью А.Д. Шмелева [12].)

Приведем пример. Исключительно популярны формы «на**й**м» и «за**й**м» при нормативных *на**ё**м* и *за**ё**м* (*госза**ё**м*). Здесь наблюдается влияние косвенных (более употребительных) падежей и «выравнивание» парадигмы. Слова *на**й**м* и *за**й**м* настолько часто в наше время встречаются даже в речи достаточно образованных людей, что уже почти превращаются в норму (но Word'2013 пока еще держится и подчеркивает эти формы красным).

➢Разнобой в написании заимствованных слов и проблема полноты системного словаря

В последние два-три десятилетия в русский язык хлынули многочисленные заимствования – в основном американского происхождения. Орфография таких новых слов еще не устоялась: в словарях они появляются, как правило, со значительной задержкой. В текстах можно встретить разные варианты написания заимствований из английского языка: «бэби» вместо нормативного беби [1; 3; 9] (вордовский спеллер, наоборот, правильным считает часто встречающееся «бэби»); «тинэйджер» вместо тинейджер [1–3; 9] (но в системном словаре-2013 есть только «тинэйджер»); «хэппи-энд» вместо хеппи-энд [1–3; 9] (Word признаёт оба варианта); «сканнер» вместо сканер [1–3; 9] (от англ. scanner; автокорректор признаёт оба варианта).

При перегруженности вордовских системных словарей низкочастотной и устаревшей лексикой некоторые вполне употребительные (и даже не очень новые) слова в нем отсутствуют.

Приведем примеры слов из «Русского орфографического словаря» [9], подчеркиваемых спеллером Word'а как неопознанные: *бивалютный, блогер, воспроизводимость, гламур* (подсказка-

2013 среди прочих вариантов выдает словоформу «глуму»), *коллайдер* («адронный» тоже подчеркивается), *комплексовать, конфискат, кремлинолог, мониторить, откудова* (можно не держать это просторечное слово в системном словаре, но подсказка-2013 предлагает «оттудова»!), *наркодилер* (подсказка предлагает «народили»), *нестыковочка, переусложнять, погрузоразгрузка* ([2], но такого слова нет в словарях [3; 9]), *подредактировать, политкорректный* (по версии спеллера-2013: «корректный» чем-то «полит»!), *рефлексивность, сверхрадикальный, соинвестор, спецсеминар, суперЭВМ* (подсказка-2013 выдает «суперэвм» и «супер-эвм»), *телегеничность, тыща* (подсказкой предлагается «тёща», но не «тысяча» – много букв!), *унисекс* (подсказка-2013 выдает словоформу «унисексов»). Этот список пока не внесенных в системные словари-2013 слов, разумеется, далеко не полон.

В Ворде-2013 по неизвестным причинам подчеркиваются красным как неопознанные следующие словоформы: *дублирующаяся, дублирующееся, дублирующиеся* (подсказка-2013 предлагает *дублирующ**ий**ся*), *минимизирующий* (хотя глагол *минимизирова**т**ь* в системном словаре-2013 имеется), краткие формы *релевантен*, *релевантно, релевантна, релевантны* (хотя прилагательное *релевантный* пропускается без подчеркивания).

Выводы

Из всего вышеизложенного можно сделать, в частности, следующие выводы. Желательно, чтобы в системных словарях текстового редактора выявлялись порождаемые низкочастотные словоформы, которые могут **совпасть с искажениями в результате наиболее вероятных ошибок и опечаток** достаточно употребительных словоформ. Такие «подводные камни», которые препятствуют эффективному выявлению ошибок (тем более при недостаточно разработанном Word'овском анализе синтаксического строения фраз), можно либо заблокировать, либо снабжать особыми пометами – предупреждениями для пользователя.

При разработке очередных версий автокорректоров (спеллеров) желательно изучать статистику употребления словоформ в текстах, а также статистику ошибочного их написания. При этом полезно учитывать причины – технические и психологические – происхождения опечаток и ошибок для их прогнозирования.

Мы предлагаем дополнять компьютерные системные словари списками наиболее вероятных искажений, в том числе не только однобуквенных, с их исправлениями. Например, подобный список мог бы состоять из пар вроде {**тыща, тысяча*}, {**тыщи, тысячи*}, {**болезн**енен**, *болезнен*}, {**д**ос**видани**е**, до свидани**я***} и т.п. Такие перечни типичных и «популярных» ошибок будут способствовать более эффективной работе автокорректора и минимизации числа отказов в выдаче вариантов исправления при компьютерной проверке текстов.

Желательно, чтобы сервисная программа-подсказка ранжировала найденные «похожие» на неопознанное слово словоформы, выдавая первыми наиболее вероятные варианты его исправления. Списки «популярных» искажений слов и словосочетаний могут помочь в выполнении и этой задачи.

Игорь Станиславович Ашманов в 2009 году писал: «Я своими руками сделал русскую морфологию в ОРФО много лет назад <…> Короче говоря, улучшать спеллеры можно. Но это вряд ли окупится, если не будет гранта или госфинансирования. Потому что продать пользователям следующую версию спеллера, если в нем есть тончайшие улучшения типа "меньше стали путаться редкие слова и ошибки" – нельзя» (https://roem.ru/27-03-2009/128340/yandeks-poka-ne-budet-delat-brauzer/). Увы!..

Литература

1. Большой иллюстрированный словарь иностранных слов: 17 000 сл. — М.: «Изд-во АСТ», «Астрель», «Русские словари», 2002. – 960 с.
2. Букчина Б.З., Сазонова И.К., Чельцова Л.К. Орфографический словарь русского языка. – 4-е изд., испр. – М.: АСТ-ПРЕСС КНИГА, 2008. – 1296 с. – (Настольные словари русского языка).
3. Зализняк А.А. Грамматический словарь русского языка: Словоизменение. Ок. 110 000 слов. – Изд. 6-е, стер. – М.: АКТ-ПРЕСС КНИГА, 2010. – 800 с. – (Фундаментальные словари).
4. Казакевич О.А., Членова С.Ф. Полвека лаборатории автоматизированных лексикографических систем НИВЦ МГУ им. М.В. Ломоносова // Вестник Российского государственного гуманитарного университета. – 2014. – Т. 16, № 8. – С. 28–39.
5. Лавошникова Э.К. Word: Причины пропуска ошибок при компьютерной проверке правописания // Science Time. – 2015. – № 6. – С. 271–275.
6. Лавошникова Э.К. О компьютерной проверке синтаксических конструкций в текстах на русском языке // Информационные процессы. – 2005. – Т. 5, № 3. – С. 201–212.
7. Лавошникова Э.К. Фамилии, имена, отчества и текстовый редактор MS WORD // Science Time. – 2015. – № 8 (20). – С. 93–99.

8. Лавошникова Э.К. О компьютерной коррекции «популярных» ошибок в текстах на русском языке // Научно-техническая информация. Серия 2. «Информационные процессы и системы». – 2003. – № 9. – С. 28–34.
9. Русский орфографический словарь (РОС): ок. 200 000 слов / под ред. В.В. Лопатина, О.Е. Ивановой. – Ин-т русского языка им. В.В. Виноградова РАН. – М.: АСТ-ПРЕСС КНИГА, 2013. – 896 с.
10. Спира И. Microsoft Excel и Word 2013: учиться никогда не поздно. – СПб: Питер, 2014. – 256 с.
11. Успенский В.А. Субъективные заметки о неправильной норме // Русский язык сегодня. Вып. 4. Проблемы языковой нормы. Сб. статей / Ин-т рус. яз. им. В.В. Виноградова РАН – М., 2006. –С. 537–571.
12. Шмелев А.Д. Распространенная ошибка или новая норма: как отличить одно от другого? // Отечественные записки. – 2014. – № 2 (59) – С. 274–285.