

КОМПЬЮТЕРНАЯ КОРРЕКЦИЯ ОРФОГРАФИИ И РАЗНОЧТЕНИЯ В СЛОВАРЯХ: ВОЗМОЖНЫ ВАРИАНТЫ?

COMPUTER CORRECTION OF SPELLING AND DIFFERENCES IN DICTIONARIES: POSSIBLE OPTIONS?

Рассматривается проблематика компьютерных систем проверки правописания. Работа автокорректоров разбирается на примере ОРФО, встроенного в текстовый редактор MICROSOFT WORD. Приводятся примеры расхождений в рекомендациях некоторых авторитетных словарей русского языка и примеры слов с «неоднозначной» орфографией, что особенно характерно для заимствований. Исследуются психологические причины написания слов с ошибками. Даются рекомендации для разработчиков новых версий спеллеров с программами-«подсказками».

Ключевые слова: проверка правописания, орфографические ошибки, психология ошибок, орфографический словарь, спеллер, MS Word.

Computerized proofing issues are discussed. Automatic corrector's functionality is exemplified on the popular embedded spelling tool ORFO that can be found inside a text editor MICROSOFT WORD. The article shows examples of doubtful recommendations of some popular dictionaries with ambiguous spelling, particularly for loanwords. Psychological reasons for spelling errors are studied. Author gives recommendations on how to develop new versions of automatic spelling correction software.

Keywords: spell check, spelling errors, psychology errors, spelling dictionary, speller, MS Word.

Одна из самых распространенных в настоящее время систем проверки правописания – автокорректор ОРФО, встроенный в текстовый редактор Microsoft Office Word 2003, на который мы и будем ссылаться. При компьютерном наборе авторы текстов, как правило, полагаются на рекомендации этого спеллера, которые не всегда безупречны даже в отношении орфографии: синтаксиса в этой статье мы не касаемся, но ошибочных замечаний Word'овского спел-чекера к построению проверяемых фраз гораздо больше [6].

При проверке орфографии спеллер сравнивает очередное слово со словоформами, имеющимися или порождаемыми (из специальным образом сконструированных квазиоснов и квазиокончаний) в словарях автокорректора [5]. Если слово в системных словарях не найдено, то оно либо «новое», т. е. пока еще не включено в словари текстового редактора, либо содержит ошибку.

Примеры словоформ, *не порождаемых в системных словарях Word'овского текстового редактора и подчеркиваемых в проверяемых текстах красной волнистой линией*, мы тоже будем подчеркивать.

В дневниках 1963 года А.К. Гладкова [8] встречаются слова в «старорежимном» написании: *матерьял* (есть в словаре Д.Н. Ушакова 1935–1940 гг. [10] наряду с *материал*), *чорт* [10], а также «логичное», но неправильное *оперетточный*. В современном «Русском орфографическом словаре» [9] имеются слова: *материал* (и *матерьял* как вариант «при передаче произношения, в поэзии»); *чёрт* (без вариантов); *оперетточный*, *оперетка*, а также *антенка* без удвоения согласной; но в уменьшительном деривате *программка* удвоение «м» сохраняется.

Некоторые варианты написания слов, отсутствующие в словарях [1; 3; 9], часто встречаются в текстах первой половины XX века. Примеры из словаря Ушакова [10]: *варьянт* (наряду с *вариант*), *жолоб* (и *жёлоб*), *шкап* (и *шкаф*). Спеллер вордовского текстового редактора считает ошибочным рекомендуемое современными словарями написание *мелочовка* [3, 9] и предлагает при вызове программы-«подсказки» бывшее не так давно нормативным *мелочёвка*; при проверке не подчеркивается ни *грегорианский*, ни *григорианский* (словари [1; 3; 9] рекомендуют без вариантов второе написание, но в [10] были даны оба варианта).

Если расхождения со словарем второй половины 1930-х гг. вполне объяснимы, то разночтения в словарях, выпущенных солидными издательствами большими тиражами уже в XXI веке и практически в одно и то же время, вызывают вопросы. При сравнении мы видим, что орфография некоторых (и не только заимствованных) слов даже в авторитетных современных словарях разнится. **Примеры:**

видеоплеер [1; 3; 9] и *видеоплейер* [2] (спеллер ОРФО пропускает без подчеркиваний оба варианта; ср. со «старым» заимствованием *конвейер*);
горнолыжный [3; 9] и *горно-лыжный* [2];
дистрибьютор [3; 9], *дистрибутор* [9] и *дистрибьютер* [1];
естественно-научный [2; 9] и *естественнонаучный* [3];
койко-место [9] и *койкоместо* [2];
капучино (кофе) [9] и *капуччино* [1];
киднеппинг [9] и *киднепинг* [1; 3];
массмедиа [1; 9] и *масс-медиа* [2];
онлайн [9] и *он-лайн* [1; 2] (нет подчеркнутых слов в Word'овских системных словарях);

погрузо-разгрузочный [9] и *погрузо**ра**згрузочный* [2] (спеллер признает оба варианта);

попкорн [9] и *поп-корн машина* [2];
роуминг [1; 9] и *роминг* [1];
секонд-хенд [9] и *секонд-хэнд* [1];
уик-энд [2; 9] и *уикенд* [1; 3];
фитнес-клуб [9] и *фитнесс-клуб* [1];
форс-мажор [1; 2; 9] и *форс**ма**жор* [3];
хеви-метал [1, 9] и *хэви-металл* [2];
человеко-час [3; 9] и *человеко**ч**ас* [2];

членкор [2; 3; 9] и *чл.-корр.* [2; 9], но не *член-корр.* (Word'овский автокорректор признает все три варианта);

чучхе (корейск.) [9] и *чучхэ* [1].

При таких несоответствиях в современных авторитетных словарях напрашивается вопрос: **а что же считать правильным написанием слова?**

Справедливости ради отметим, что в последние десятилетия в лексику русского языка вошло много заимствований – в основном американского происхождения [11, 13].

Орфография таких слов еще не устоялась, а в словарях они появляются, как правило, значительно позже, чем в текстах и устной речи, поэтому и в текстах тоже встречается много вариантов написания заимствованных слов.

В текстах можно встретить такие варианты написания заимствований из английского языка:

бойфрэнд или *бой-френд* (вариант спеллера) вместо *бойфренд* [1; 3; 9];

бэби вместо *беби* [1; 3; 9];

кэш (весьма распространенный в текстах вариант спеллера, что отмечается также в статье [11]) вместо *кеш* [9];

лэйбл вместо *лейбл* [1, 9];

мьюзикл и *мьюзик-холл* вместо *мюзикл* [1; 3; 9] и *мюзик-холл* [1–3; 9];

оффис вместо *офис* [1; 3; 9] (англ. *office*; удвоение согласной в заимствованиях по традиции в нормативном написании часто снимается, ср. со «старым» заимствованием *офицер*);

оффшорный вместо *офшорный* [1–3; 9] (от англ. *off shore*);

риэлтор (вариант спеллера), *риэлтер* (на это написание вордовская «подсказка» дает только *риэлтор*) или *риелтер* (а в этом случае выдается только *риелтор*) вместо *риелтор* [1; 3; 9] (англ. *realtor*);

сканнер вместо *сканер* [1; 3; 9] (англ. *scanner*; спеллер признает оба варианта);

тинэйджер вместо *тинейджер* [1; 3; 9] (но в системном вордовском словаре только *тинэйджер*);

хэппи-энд вместо *хеппи-энд* [1; 3; 9] (спеллер признает оба варианта).

Вот некоторые примеры дериватов, отсутствующих в системных словарях текстового редактора MS Word: *лидерствовать*, *невыездной*, *нетипичность*, *продвинутость*, *прокремлёвский*, *псевдорадикальный*, *рокировочка*, *самопиар* и т. п.

Желательно подобные слова какими-либо способами порождать от уже имеющихся основ. То есть усложнение механизмов **словообразования** – задача для будущих версий автокорректоров.

С некоторых пор вордовский спеллер перестал «узнавать» (из-за новейших коннотаций) слова *голубой* и *розовый*, а также слово *негр*. Нет в системных словарях и других «нехороших» слов (*сдохнуть*, *дерьмо* и т. п.). Они подчеркиваются красным, хотя можно было бы «понимать» эти слова и подчеркивать зеленым, как подчеркиваются просторечные слова. Многие бранные слова в ОРФО тоже отсутствуют, даже *осёл* вычеркнут из вордовского словаря, несмотря на то что *осел* (но «е» и «ё» спеллер не различает) может быть формой глагола *осесть* (формы *осела*, *осело*, *осели* тоже почему-то подчеркиваются красным). Как тут не вспомнить замечание Фаины Раневской: кое-что есть, а слова такого нет!

Приведем еще некоторые примеры рекомендаций вордовской программы-«подсказки»: для «популярного», но ошибочного **скурпулезный* выдается правильное *скрупулезный*, для *ньюйоркский* – *нью-йоркский* [2; 3; 9], но для **досчатый* «подсказка» не дает *дошчатый* [3; 9], так как в этом случае следовало бы две буквы заменить одной (другой) буквой, а выдает сообщение «Варианты отсутствуют».

Для местоимения *ихний* (прост. к *их*) [3; 9], которое система ОРФО тоже объявляет просторечным (с подчеркиванием зеленой волнистой линией), но сначала подчеркивает красным, «подсказка» должна была бы предложить нормативное притяжательное местоимение *их*, однако она выдает только один вариант – «более похожий» термин *ионий*.

Можно предположить возможные *психологические причины* возникновения наиболее типичных ошибок при порождении текстов [4] (здесь мы не говорим об опечатках). Есть стремление осмыслить и подправить непривычную морфологию услышанного слова, сделать его «более понятным».

Например, встречается написание **невропатолог* – говорящие и пишущие производят его от словоформы *нервы*, но правильно: *невропатолог* от *neuron*. При этом, однако, в словарях есть не только слово *невроз* (болезнь), но и *нервоз* (‘нервное состояние’) [3; 9]. Если в слове две ошибки (**нервопаталогия*), то программа-«подсказка» вариантов исправления, как правило, дать не может.

Примеры слов из «Русского орфографического словаря» [9] с их «популярными» искажениями пишущими:

ажитация (франц. *agitation*; но иногда происходит смешение со словом *ажисотаж* и получается **ажисотация*);

апробировать (‘проверив, одобрить’; неправильное **опробировать* – как *опробовать*);

бессребреник (встречается написание с ошибками сразу в двух местах **бессеребреник*; а в словаре Ушакова [10] другой вариант *бессребренник* – единственный);

военачальник, солдафон (иногда пишут и говорят **военноначальник, *солдатофон*: так «понятнее»);

времечко, темечко, семечко (иногда пишут через «я», так как образовано от слов на «-я»; с другой стороны,

**времяпровождение* и **времяпрепровождение* вместо нормативных *времяпровождение* и *времяпрепровождение* тоже можно встретить);

двуглавый, двугорбый (иногда пишут и произносят с «х», так как чаще встречаются слова с начальным «двух-»);

джентльмен, мужеложство, пригоришня, учреждение (говорят и пишут **джентельмен, *мужеложество, *пригорошня, *учереждение*: некоторые сочетания согласных удобнее произносить с включением гласной);

интроверт (лат. *intro* – внутрь, *vertere* – обращать [1], но вспоминают слово *экстраверт*, вместо «экс-» подставляют «ин-», и получается **интраверт*; с другой стороны, в словарях [2; 9] имеются слова *интрамолекулярный, интразональный*: они происходят от латинского *intra* – внутри [1]);

кондоминиум (от лат. *con* ‘с, вместе’ и *dominium* ‘владение’ [1], не **кондоминимум*);

кристальный (нормативно с одной «л», хотя в слове *кристалл* удвоенная «л», но в словаре Ушакова [10] два варианта: *кристалл* и *кристал*, при этом там же имеются *кристалик* и *кристальный*);

междуусобица (иногда пишут **междуусобица*);

млекопитающее (‘питающее молоком’, но иногда говорят и пишут **млекопитающееся* – по аналогии со словом *пресмыкающееся*);

нянчиться (**нянчиться* – влияет написание слова *нянька*);

перфекционизм (**перфектиционизм* – вставляют «т», образуя это слово от *перфект*);

поскользнуться (произносят и пишут **подскользнуться*);

пфенниг (встречаются **пфеннинг* и **пфенинг*: в словаре Зализняка с обратным алфавитным порядком [3] слов на «-ниг» больше нет, а на «-нинг» – достаточно много);

радионуклиды (часто говорят и пишут **радионуклеиды*, но в словарях [1, 2, 9] (*радио*)*нуклиды*, несмотря на происхождение от «лат. *nucleus* ядро» [1]);

сногшибательный (**сногосшибательный* – пишущим кажется, что в сложных словах должна быть соединительная гласная, да и сочетание «гсш» – труднопроизносимое);

спецкор [2, 3, 9] (от *корреспондент*, поэтому иногда пишут **спецкорр*);

фальстарт [1–3, 9] (англ. *false start*; неправильное **фальшстарт* «понятнее»);

электрификация, *газификация* (влияние более привычной соединительной гласной «о»: **электрофикация*, **газофикация*);

эстрогены (**экстрагены* – искажение в результате «народной этимологии»);

юрисконсульт (происходит от «лат. *juris consultus* правовед» [1], но встречается неправильное **юристконсульт* или даже **юрист-консульт*).

Не всегда с пониманием воспринимается превращение буквы «и» в «ы» (иногда мы встречаемся с «сопротивлением» авторов текстов) в словах: *небезызвестный*, *подынтегральный*, *предыстория* и т. п. Неприятие такой метаморфозы может быть вызвано тем, что «и» остается в нормативном написании слов *межинститутский*, *сверхиысканный* (иначе получились бы буквосочетания «жы» и «хы»), в сокращении *сельхозинвентарь*, а также в написании слов с заимствованными префиксами: *супериигра*, *гиперинфляция*, *панисламизм* и т. п. (В.В. Жириновский в 2014 году предлагал букву «ы» вообще исключить из русского алфавита, а также убрать все заимствования.)

В устной речи и в текстах из-за влияния косвенных форм нередко можно встретить неправильно построенный инфинитив: **скорбить*, **бдить* **запечатлить* вместо *скорбеть*, *бдеть*, *запечатлеть*, а также *придти* (этот вариант отсутствует в словарях [3; 9], но спеллер пропускает *придти* без подчеркивания – наряду с нормативным *прийти*). Слово *ас* – «слишком короткое», поэтому иногда пишут *асс*, однако есть такое слово со значением ‘монета’, поэтому спеллер слово *асс* не подчеркивает.

В словарях ОРФО довольно много слов, которые пропускаются без подчеркивания, но с большей вероятностью могут появляться в текстах пользователя в результате опечатки или ошибки. **Примеры:**

достегать (‘закончить шитье стеганого одеяла’, от глагола *стегать*),

избабиться (стать похожим на бабу),

иступленный (карандаш),

картеж (или *картёж* – игра в карты),

нажеваться (от глагола *жевать*),
подрожать (от *дрожать*),
утоняться (делаться тоньше).

Имеются подобные примеры и среди терминов: *комплемент*, *контроллер* (аппарат) и др.

Подобные низкочастотные слова, «мешающие» проверке более употребительных слов, желательно вообще убрать из словаря автокорректора или подчеркивать их зеленым с особой пометой (подробнее см. в [7]).

1. В.А. Успенский в статье «Субъективные заметки о неправильной норме» [12, 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 [14].)

Со своей стороны, в подтверждение приведем некоторые примеры. Часто можно встретить написание **дивиденты* вместо *дивиденды* («лат. *dividendus* подлежащий разделу» [1]). Ошибочный вариант встречается едва ли не чаще, так как слов на «-ент» намного больше, чем на «-енд» (см. [3]). Исключительно популярны формы **найм* и **займ* при нормативных *наём* и *заём* (программа-«подсказка» вордовского текстового редактора выдает много вариантов исправления, но правильный вариант в обоих случаях – только десятый). Мы видим влияние более употребительных косвенных падежей и попытки «выравнивания» парадигмы. Варианты **найм* и **займ* настолько часто в наше время встречаются даже у высокообразованных людей, что уже почти превращаются в норму.

Ошибки чаще всего совершаются по аналогии с более частотными словами или в попытках порождения словоформ по более регулярным моделям. Иногда в процессе освоения новой для кого-то лексики слова искажают, пытаясь сделать их «более понятными». Но приведенные выше примеры новых заимствований чаще употребляются теми, кто в какой-то степени знаком с английским языком. Поэтому для такой лексики слабее выражена тенденция к «упрощению», русификации, свойственная менее образованным слоям населения. Напротив, знание языка происхождения проявляется либо в побуквенном отображении (транслитерации), либо в приблизительной передаче звучания таких заимствованных слов – иногда при игнорировании наших языковых традиций. (Подробнее см. в работе [11].)

К сожалению, в издаваемых словарях новые словообразования и заимствования появляются со значительной временной задержкой.

2. Стремительный прогресс в компьютерной сфере практически уже снимает ограничения, связанные с объемами баз данных. С одной стороны, желательно, чтобы в каждой очередной версии системные словари текстовых редакторов как можно больше пополнялись. При этом могли бы использоваться и механизмы словообразования, создания «словарных гнезд».

С другой стороны, в основном словаре должны быть выявлены малоупотребительные слова и словоформы, которые могут **совпасть с искажениями** (в результате наиболее вероятных ошибок и опечаток) достаточно употребительных словоформ. Такие «подводные камни» способствуют пропуску слов с ошибками или «вылезают» в списках рекомендуемых вариантов исправления. Если не представляется целесообразным убрать их совсем, то они могли бы снабжаться особыми пометами для сообщений «программы»-подсказки или выноситься в дополнительные компьютерные словари для проверки текстов особой тематики [7].

Мы предлагаем дополнять компьютерные системные словари spellера списками наиболее вероятных искажений (для некоторых слов). Такие сведения о типичных ошибках будут способствовать более эффективной работе текстового редактора и минимизации числа отказов в выдаче вариантов исправления при компьютерной коррекции текстов. Желательно также, чтобы «подсказка» ранжировала варианты исправления неопознанного слова, выдавая первыми наиболее вероятные. Предлагаемые нами списки «популярных» искажений при наборе текстов могут способствовать выполнению этих задач.

Игорь Станиславович Ашманов в 2009 г. писал: «Я своими руками сделал русскую морфологию в ОРФО много лет назад <...> Короче говоря, улучшать spellеры можно. Но это вряд ли окупится, если не будет гранта или госфинансирования. Потому что продать пользователям следующую версию spellера, 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. М., 2002.
2. Букчина Б.З., Калакуцкая Л.П. Слитно или раздельно? Орфографический словарь-справочник: свыше 107 000 слов. М., 1998.
3. Зализняк А.А. Грамматический словарь русского языка: Словоизменение. Ок. 110 000 слов. 4-е изд., испр. и доп. М., 2003.
4. Лавошникова Э.К. Маргинальное в орфографии текстов на русском языке // Международная конференция «Маргиналии-2014: границы культуры и текста». Елец, 5–7 сентября 2014 г. / Ред.: М.Ю. Михеев, Б.П. Иванюк, А.Г. Кравецкий. М., 2014. С. 55–58. (<http://uni-persona.srcc.msu.ru/site/conf/marginalii-2014/thesis.htm>)
5. Лавошникова Э.К. Компьютерная проверка орфографии: вчера, сегодня, завтра // Вестник Московского ун-та. Серия 9: «Филология». 2003, № 5. С. 53–68.
6. Лавошникова Э.К. О компьютерной проверке синтаксических конструкций в текстах на русском языке // Информационные процессы. Т. 5, № 3, 2005. С. 201–212.

7. Лавошникова Э.К. О «подводных камнях» в компьютерных системах проверки правописания // Вестник Московского ун-та. Серия 9: «Филология». 2002, № 6. С. 151–163.
8. Михеев М.Ю. Александр Константинович Гладков. Дневник 1963 // Филоlogos. Вып. 22 (3), 2014. С. 78–100.
9. Русский орфографический словарь: ок. 180 000 слов (под ред. В.В. Лопатина) / Ин-т русского языка им. В.В. Виноградова РАН. 2-е изд., испр. и доп. М., 2005.
10. Толковый словарь русского языка: в 4-х т. / под ред. Д.Н. Ушакова. М., 1935–1940.
11. Турко У.И. О вариантах новых заимствованных слов // Филоlogos. Вып. 22 (3), 2014. С. 66–70.
12. 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.
13. Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А. Статистический словарь языка русской газеты (1990-е годы). Т. 1. М., 2008.
14. Шмелев А.Д. Распространенная ошибка или новая норма: как отличить одно от другого? // Отечественные записки. № 2 (59), 2014. С. 274–285.