АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'322.2

Э. К. Лавошникова

О компьютерной коррекции "популярных" ошибок в текстах на русском языке

Рассматриваются проблемы, возникающие при работе со спеллерами — компьютерными системами проверки правописания. Перечислены наиболее типичные ошибки, встречающиеся в текстах на русском языке, для многих ошибок приведены возможные психологические причины их возникновения. Работа автокорректора разбирается на примере самого распространенного — ОРФО, встроенного в текстовый редактор MICROSOFT WORD. Даются рекомендации для разработчиков новых версий автокорректоров, поскольку сервисная подсказка часто выдает слишком много вариантов исправления, а для слов с неоднобуквенными ошибками, как правило, ничего предложить не может. Высказано предложение — дополнять внутренние словари системы списками наиболее "популярных" искаженных слов.

Автоматические или автоматизированные корректоры (спел-чекеры, спеллеры) пока не могут решать все задачи по проверке текста, так как ошибки бывают не только орфографическими или синтаксическими, но и смысловыми, логическими, фактическими, стилистическими и др.

В статье будут рассматриваться наиболее типичные орфографические ошибки в текстах на русском языке и особенности компьютерной коррекции ошибок разного рода. При этом мы будем ссылаться на один из самых распространенных спеллеров — автокорректор ОРФО, встроенный в текстовый редактор MICROSOFT WORD2000 (в этой версии используется тот же автокорректор, что и в версии 1997 г.).

Приведем пример смысловой ошибки. Некая радиостанция объявила: *Бомбовый удар был нанесен в шесть утра по Москве.* Имелось в виду — "по московскому времени".

Наиболее типичная ошибка иностранцев, даже тех, кто успешно освоил русский язык, — неразличение совершенного и несовершенного вида глаголов. Например: "Он часто *приехал* в Москву", "Я буду вам *написать*". Русскоязычный человек никогда таких ошибок не сделает, разве что при небрежной правке текста.

В "Грамматическом словаре русского языка" А. А. Зализняка [3] каждый глагол имеет пометуказание, к какому виду он относится ("св" или "нсв"). Автокорректор ОРФО, построенный, как и некоторые другие спеллеры, на этом словаре, похоже, на неправильное употребление совершенного или несовершенного вида не реагирует.

Далее мы будем употреблять термин **словоформа**, который означает конкретное слово в одной из его грамматических форм. Например, слова *идти, идет, шел, шла, идя, идущий, шедший* и т. д. являются разными словоформами **лексемы** идти. Формально можно считать, что словоформа — это просто цепочка букв в тексте между двумя ограничителями (пробелами, знаками препинания и т. п.).

Орфографические ошибки и опечатки спел-чекерами чаще всего выявляются методом поиска в их внутренних словарях каждой словоформы текста.

1. ПОДСКАЗКА — ТОЛЬКО ДЛЯ ОДНОБУКВЕННЫХ ОШИБОК?

В автокорректоре ОРФО для каждого **неопознанного** слова, т. е. отсутствующего в его внутренних словарях, **сервисная программа-подсказка** в случае ее вызова пользователем пробует заменить каждую букву другой буквой, убрать одну букву или дефис, поменять местами две смежные буквы, вставить одну букву, дефис или пробел[1] (но почему-то не тогда, когда два слова разделены запятой или точкой без пробела). Подсказка пробует убрать пробел между неопознанным словом и предыдущей, а также следующей словоформой (склеить две словоформы). Полученные такими способами "слова" подсказка пытается найти во внутренних словарях системы. Все, что найдено, выводится в список предлагаемых вариантов исправления данного слова.

Справедливости ради следует отметить, что подсказка ОРФО выдает варианты исправления не только однобуквенных, но и некоторых двойных ошибок: для искаженных слов **коммисар*, **комунна*, **имунnодефицит* или **иммунодифецит* предлагает правильное написание *комиссар, коммуна, иммунодефицит*. Т. е. подсказка может удвоить букву и убрать удвоение другой буквы одновременно в двух местах некоторых слов, а также переставить расположенные через одну буквы.

[1] Например, для ненайденного слова **сосиськи* вместе с правильным вариантом *сосиски* подсказка ОРФО выдает сочетание "*со сиськи*". Однако при отсутствии пробела после частицы *не* во фразах "Он *ненамерен* жениться" и "Они *необязаны* все знать" ОРФО в первом случае предлагает единственный вариант исправления *неманерен*, а во втором — выдается сообщение "варианты отсутствуют".

Однако для искаженных слов **буЛгаХтер* или **землЯтрЕсение*, в которых буквы, разделенные двумя буквами, поменялись местами, ОРФО правильных вариантов *бухгалтер, землетрясение* не предлагает.

Иногда в речи и на письме вставляется лишний слог: **воеНОначальник* или **воеННОначальник* (правильно — *военачальник*), **деморализИРовать (деморализовать)*, **увяНУл (увял)*, **разбухНУВший (разбухший)*, "двое *румынов*", "пара *сапогов*" (по регулярной модели, но правильно — "двое *румын*", "пара *сапог*"). Во внутреннем словаре ОРФО есть устаревшее слово "млеко", поэтому подсказка разбивает ошибочное слово **млекопитающееся* на два — "млеко питающееся", но не предлагает правильного *млекопитающее*. Пользователь может решить, что привычное для него слово просто отсутствует в словаре спеллера.

Память и быстродействие современных компьютеров уже позволяют снимать те ограничения, которые имелись раньше, поэтому можно было бы расширить количество проверок — исследовать гипотезы о неоднобуквенных искажениях. Однако если неопознанные слова подвергать многобуквенным заменам, то будет получаться слишком много неадекватных вариантов исправления.

Желательно, чтобы разработчики при дальнейших усовершенствованиях автокорректоров включали в систему списки наиболее часто встречающихся искаженных словоформ с их исправлениями.[2]

В этот список следовало бы **внести и слова с однобуквенными ошибками**, поскольку в них могут появиться другие ошибки. Если в слове *компьютер* сделать две ошибки **компЪютОр*, то подсказка ОРФО ничем помочь не сможет. Встречается написание **нерВопатолог* ("народная этимология" — не от *neuron* (см. [1]), а от *нерв*, вполне логично!). Если в таком слове будет **допущена еще одна ошибка**, то подсказка ОРФО уже ничего предложить не сможет. Однако если пары {"нервопатолог", *невропатолог*}, {"нервопатолога", *невропатолога*} и т. д. в новых версиях автокорректора занести в особый список, то при появлении в тексте слова с ошибками сразу в двух местах, например *нерВопОтологу*, после очередной замены — второй буквы *о* буквой *а* — полученное слово будет найдено в этом списке, а подсказка сможет выдать второй элемент данной пары — правильную словоформу *невропатологу*.

Для составления **списка популярных искажений** следует накапливать статистику того, какие слова чаще остаются неопознанными. Это полезно и для пополнения словарей автокорректора новой лексикой.

Если бы в такой список были внесены пары с неоднобуквенными искажениями {"пресмыкающее", *пресмыкающееся*}, {"болезненен", *болезнен*} и т. п., то подсказка была бы более эффективной.

Узкоспециальные термины вряд ли попадут в этот перечень "популярных" искажений, так как специалисты знают свою терминологию и допускают скорее опечатки, чем ошибки. Однако если термин становится широкоупотребительным, то он может иногда "адаптироваться" и искажаться.

2. НАМЕРЕННЫЕ ОШИБКИ И СЛОВЕСНАЯ ИГРА

Ошибки обычно бывают непроизвольными, но бывают и намеренными.

Намеренные ошибки и искажения пользователь исправлять не будет. Рассмотрим следующие примеры: "Глубокоуважаемый *вагоноуважатый*, <...> нельзя ли у *трамвала вокзай* остановить?" — обращался к вагоновожатому "человек рассеянный" в стихотворении С. Я. Маршака. В данном случае намеренные "ошибки" в тексте продуцирует автор, а не герой произведения.

Намеренное искажение слов мы видим в художественной литературе при авторской передаче акцента, недостатков произношения — шепелявости, картавости и т. п. Встречается отображение неправильной детской речи.

В художественных текстах может воспроизводиться малограмотная речь персонажей. Приведем примеры из песен Высоцкого: "он *пройтиться* хотел по нейтральной земле" (здесь стилизованная авторская речь), "*тута* есть культурный парк". Для просторечного *пройтиться*, которое подчеркивается как неопознанное, подсказка ОРФО выдает единственный вариант "исправления" — *проститься*. Слово *тута* воспринимается спеллером как существительное (означает "тутовое дерево" — см. [3]) и пропускается без замечаний.

К намеренным искажениям можно отнести некоторые примеры языковой игры: "оборзение" вместо *обозрение*, "засланец", "нервомотолог", "сексплуатация", "нуворишки", "мы непокобелимы". Реклама пива призывает нас "жить припИваючи". В перестроечные времена наряду с биржами появлялись и "буржи", как их называли создатели таких заведений.

Слово "хрущоба" (контаминация фамилии *Хрущев* и слова *трущоба*) уже вошло в "Русский орфографический словарь" [10].

Ценные бумаги иногда защищают от подделок с помощью намеренных ошибок в микротексте. "Ошибки" при этом не должны быть грубыми[3] и очевидными.

Изучение английского языка иногда приводит к тому, что мы видим написание "траффик" вместо *трафик* (см. [1], [10]) и т. п. Возможно, не все и не во всех случаях согласны подчиняться авторитету академических словарей.

3. НАИБОЛЕЕ ПОПУЛЯРНЫЕ ОРФОГРАФИЧЕСКИЕ ОШИБКИ

Непроизвольные ошибки можно разделить на **ошибки правописания**, происходящие от недостаточного знания орфографии и грамматики, и **опечатки**. Более подробно об опечатках — в нашей статье [7]. Нередко тот, кто уличен в недостаточной грамотности, пытается выдать свои ошибки за простые опечатки (якобы по невнимательности). Действительно, четкую границу здесь трудно провести.

[2] В MS WORD есть сервисная команда Автозамена. Желательно, чтобы пользователь сам составил список своих "излюбленных" ошибок и опечаток.

[3] В восприятии человека ошибки могут быть более грубыми и менее грубыми. Наверное, в написании **аклематизация (акклиматизация)* первая ошибка менее грубая, чем вторая, а в написании **згущонка (сгущенка)* — наоборот.

Характер и частота опечаток в значительной степени зависят от *устройства клавиатуры* и **другой компьютерной специфики** (при сканировании, например). Подсказка часто предлагает длинный перечень вариантов исправления, не упорядоченный по их вероятности. Можно было бы в первую очередь выдавать словоформы, *уже имеющиеся в тексте.*

Орфографические ошибки. Ниже приводятся примеры орфографических ошибок, *встречающихся даже у сравнительно грамотных людей.* Думается, что перечень таких ошибок *вполне обозрим.* Подобная информация может быть внесена разработчиками в новые версии автокорректоров в виде пар {искаженное слово, его правильное написание}"[4], для того чтобы не искать и не выдавать в подсказке маловероятные варианты исправления. При появлении в слове предусмотренной ошибки и еще одной ошибки или опечатки будет сохранена возможность выдачи правильного варианта.

Иногда вставляется *лишняя согласная* в слова:

дерматин, желатин (говорят и пишут **дермаНтин*, **желаНтин*), *инцидент, прецедент*;

компрометировать (вставляют лишнюю букву "н", но зато пишут **трансцедентный* вместо *трансцеНдентный*);

констатировать (от лат. *constat* — известно, **констаНтировать* производят от слова *константа?*);

конкурентоспособный (**конкурентНоспособный*), пфенниг (**пфенниНг* — слов, оканчивающихся на "-ниг", в словаре А. А. Зализняка [3] с обратным алфавитным порядком больше нет);

участвовать (**учаВствовать* — под влиянием глагола *чувствовать?*);

яства (**яВства* — под влиянием прилагательного *явственный?*).

Неправильное произношение и написание делает следующие слова психологически "более понятными": **грейпфруКт* или даже **грейфрукт* вместо *грейпфрут*, **двуХглавый* (правильно *двуглавый*, но *двухголовый*), **поДскользнуться* (*поскользнуться*), **суБпрефект*[5], **фальШстарт* (*фальстарт*).

Как разновидность вставки лишней буквы встречается ошибочное **удвоение согласной** в словах:

ветреный (день, человек);

гостиница, гостиная;

грамотный (**граММотный* — под влиянием слова *грамматика*); дилер (**диЛЛер* — как *киллер?*);

директриса (**директриССа* — как *клоунесса*);

импресарио (**импреССарио* — наверное, с удвоенной "с" слово выглядит более солидным и иностранным);

мороженое или *свежемороженый* (но "мясо, долго *мороженное*" — в этом сочетании требуется двойное "н");

оперетка (**опереТТка*, обратный случай — пишут **програмка* вместо *программка*);

путаница, свояченица;

расчетливый, расчет (**раССчет* — под влиянием глагола *рассчитать*);

серебряный;

стела (**стеЛЛа* — под влиянием слова *стеллажи* или имени *Стелла?*);

труженик (**труженНик*, как *презренник?*);

юность, юный (**юНный* под влиянием сокращения *юннат*, т. е. "юный натуралист");

в сокращениях *групорг* (**груППорг*), *партгрупорг, профгрупорг.*

В следующих словах в речи и на письме иногда появляется *лишняя гласная*: *будущий* (по аналогии со словом *следующий* получается **будуЮщий*), *мужеложство, перспектива, перетурбации* (**переспектива*, **перетурбации* — приставка "пере-"?), *пригоршня* (**пригорОшня*), *сногсшибательный* (**сногОсшибательный* — некоторым кажется, что в сложных словах обязательно должна быть соединительная гласная), *учреждение* (**учЕреждение*), *чрезвычайный.* Нередко говорят и пишут **радионуклЕиды* (в [1, 2, 10] (*радио*)*нуклиды*, несмотря на происхождение от лат. *nucleus* — ядро).

В других случаях, напротив, *выпадает гласная*, находящаяся в слабой позиции, в словах: *заведуЮщий, канцЕрогенный, папорОтник, притолОка, судорОга, сутолОка* и др.

В речи и на письме можно встретить ошибки в словах: *гауптвахта* (**гауптвахта*), *дивиденды* (**дивиденТы* — встречается едва ли не чаще, слов на "-ент" намного больше, чем на "-енд"), *матерщинник* (**матерШинник*), *скрупулезный* (**скурпулезный*), *фольклор* (**фольклЁр*), *хахаль* (**хахЕль*), *электрификация, газификация* (**электрОфикация*, **газОфикация*).

Каждое из приведенных выше слов с однобуквенными ошибками или с перестановкой смежных букв ОРФО подчеркивает красной волнистой линией и дает в подсказке правильный вариант исправления, но часто среди нескольких вариантов и не первым.

В процессе превращения в более широко употребляемые слова иногда не обходятся и без искажений, в том числе и неоднобуквенных: *аккредитив, апелляция, аппликация, аттракцион, бессребреник* (**бессЕребреНник* — от полногласного *сЕребро* и с двумя *н*), *диссидент, индифферентный, инжиниринг* (англ. *engineering*, но *инженер*), *интеллигент* (нередко пишут **интеллЕгент* под влиянием слова *интеллект*), *периферия* (**перЕферия* — приставка "пере-" имеется в виду?), *превалировать, привилегия, рентген, ртутьсодержащий* (**ртутОсодержащий* — якобы в сложных словах соединительная гласная обязательна), *эликсир* (**элЕксир*).

В школе нас учили правильно писать слова *солнце, лестница* и др. Зачастую вызывает затруднения написание слов: *винегрет, времяпрепровождение, искусственный, печенка, семечко* (**семЯчко*), *сумасшедший* (**сумашедший* — как слышим, так и пишем), *тушенка* (от *тушить*, **тушОнка* — по аналогии со словом *душонка?*).

С трудом дается применение правила написания приставок, изначально оканчивающихся на *з*, перед глухими согласными. Компьютерщики любят заниматься словообразованием от английских

[4] Не всем известно, что литературный псевдоним знаменитого американского писателя *О. Генри* пишется через точку, а не через апостроф (см. [3]). Пару {"О'Генри", *О. Генри*} также можно было бы занести в предложенный список.

[5] Однако ОРФО считает правильными одновременно слова *супрефектура* и *суБпрефект*, а слово *супрефект* [1, 3, 8, 10] подчеркивает как неправильное.

терминов, при этом глагол *расконнектиться* пишут иногда через *з*. Зато в одной газете недавно встретилось написание **бесвкусица* (*безвкусица*).

В искажении следующих слов повинна так называемая "народная этимология": *апробировать* (проверив, одобрить, но пишут **опробировать* — как *опробовать*?), *довлеть* (преобладать, господствовать, тяготеть, но **давлеть* — как *давить*?), *кондоминиум* (от лат. *dominium* — владение, **кондоминиМум* — "кондовый минимум"?), *междоусобица* (**междуусобица*).

Мы видим, что часто ошибки происходят от уподобления слова более употребительным словам[6], от стремления сделать его "более понятным".

Не всегда легко воспринимается превращение буквы "и" в "ы" в словах: *небезызвестный, подындекс* [10], *подынтегральный* [3, 8, 10], *предыстория* и т. п. При этом буква "и" остается в нормативном написании слов *гиперинфляция* [2, 3, 10], *межинститутский* [2, 3, 8, 10], *сверхинтересный* [2, 8, 10], *панисламизм* [2, 3, 8, 10], *трансиорданский* [2, 3, 8] и т. п.

Нередко пишут "*кверх* ногами, тормашками" (правильно — *вверх*...), "ей богу" вместо дефисного написания *ей-богу* [2, 3, 8, 10], **полоборота*[7] при нормативном написании через дефис перед гласной.

Нередкая ошибка — сращение частицы *не* с глаголами *хватать, хватить*. Конструкция *не хватает* воспринимается в качестве одного слова (возможно, под влиянием глагола *недоставать*).

4. ТВЕРДЫЙ ЗНАК — НЕ ЛИШНЯЯ ЛИ БУКВА?

Некоторые слова вызывают у пишущих сомнения, нужен ли в них твердый знак: *межъязыковой, трёхъярусный, панъевропейский, фельдъегерь, артъярмарка* [10].

В то же время без твердого знака пишутся сокращения *детясли* [2], *комячейка* [2], *партячейка* [2, 10], *Минюст* и т. п.

Твердый знак вообще можно было бы исключить из русского алфавита.

Однако заменять его в таком случае имеет смысл не апострофом, как это было в первые годы советской власти, а мягким знаком, тем более что на глаз они почти неразличимы. К тому же освободилось бы место на компьютерной клавиатуре в русском регистре для других символов.

В словах *адъюнкт, адъютант, дизъюнкция, конъюнктурщик, конъюнкция* слышится смягчение согласной перед "ъ".

В некоторых словах (*арьергард, интерьер, обезьяна*) иногда пишут по ошибке твердый знак вместо мягкого знака.

Известный языковед Ф. Ф. Фортунатов уже в 1904 г. высказался за полную отмену буквы ъ.

Препятствий в современном русском языке к этому нет, поскольку отсутствует такая пара, слова в которой различались бы только наличием ъ и ь в одной и той же позиции.

В. А. Успенский в статье "Одна модель для понятия фонемы" (Вопр. языкознания. — 1964. — № 6. — С. 53) пишет: "...Если считать буквы ь и ъ графически близкими (что довольно естественно), то они окажутся, по-видимому, аллографами одной графемы в письменном литературном русском языке с обязательным употреблением буквы ё". И далее в сноске В. А. Успенский добавляет: "Если не употреблять букву ё, то возникает противопоставление *вьемся—въемся*, любезно указанное автору А. А. Зализняком. За пределами литературного языка можно найти, по-видимому, еще несколько противопоставлений..."

При исключении из русского алфавита твердого знака и замене его мягким знаком слова *въемся* и *въешься* (малоупотребительные формы глагола *въесться*) сольются в омографы со словоформами *вьёмся* и *вьёшься* — в случае написания последних через *е*, что вряд ли может служить аргументом против нашего предложения.

Впрочем, на первых порах можно было бы разрешить писать в словах, в которых есть ъ, как твердый знак, так и мягкий знак — по желанию. С другой стороны, желательно восстановить в правах букву ё.

5. ЧТО ЖЕ НАМ РЕКОМЕНДУЮТ СЛОВАРИ?

В последние годы на нас обрушилось много заимствований — в основном американизмов. Орфография таких новых слов еще не устоялась, в словарях они появляются, как правило, значительно позже, чем в текстах.

Далее в примерах слова, отсутствующие в словарях [1, 2, 3, 8] и [10], помечены [*].

Прослеживается тенденция написания английских заимствований через *"э"*, т. е. противодействие их "обрусению". В некоторых случаях словари с этим борются, но без большого успеха. В текстах может встретиться такое написание:

"бэби" [*] вместо *беби* [1, 3, 8, 10];

"лэйбл" [*] вместо *лейбл* [1, 8, 10];

"риЭлтЕр" [*] или "риЭлтор" [*], но в последних изданиях ряда авторитетных словарей узаконено написание *риелтор* [1, 3, 10], "римейк" [*] (как слышится по-английски, так и пишется) вместо рекомендуемого *ремейк* [1, 3, 10] (с употребительным в русском языке префиксом "ре-");

"хэппи-энд" [*] вместо *хеппи-энд* [1, 2, 3, 8, 10] (но *уикенд* [1, 3, 10], после дефиса и в начале слова — буква "э").

Однако справедливости ради отметим, что написание некоторых слов в последних изданиях словарей разнится. Например:

видеоплеЙер [2] и *видеоплеер* [1, 3, 10];

горнолыжный [3, 8] и *горно-лыжный* [2, 10];

зек [8] и *зэк* [3, 10];

киднеППинг [8] и *киднепинг* [1, 3, 10];

кремлЕнолог [8] и *кремлИнолог* [10];

мелочЁвка [8] и *мелочОвка* [3, 10];

погрузо-разгрузочный [8] и *погрузоразгрузочный* [2, 10];

секонд-хЭнд [1] и *секонд-хЕнд* [10];

[6] В авторской телепередаче А. Ливанской пожилая деревенская женщина сказала, что она выписывала газету "АГРУменты и факты". Можно предположить, что у нее на слуху были слова *агроном, агротехнический* и т. п.

[7] Подсказка ОРФО наряду с *пол-оборота* в качестве варианта исправления этой ошибки предлагает и такое — "полобормота".

травмпункт [8] и *травмОпункт* [3, 10] (в [2] даны оба варианта);

уик-энд [2, 8] и *уикенд* [1, 3, 10];

форс-мажор [1, 2, 8] и *форсмажор* [3, 10];

хэви-металл [2] и *хеви-метал* [1, 10];

человеко-час [3, 8] и *человекочас* [2, 10].

Возникает вопрос, а что же считать правильным написанием слова?

6. БОЛЬШЕ СТИЛИСТИЧЕСКИХ ПОМЕТ!

С одной стороны, в словарях не всегда можно найти новые слова, с другой — они бывают перегружены устаревшей лексикой, причем часто без каких-либо помет.

Как было показано в нашей статье [6], во внутреннем словаре автокорректора желательно свести к минимуму количество "подводных камней", т. е. теоретически возможных, но практически не употребляемых словоформ. Например, слова *брег, ветр, огнь, угль*, не имеющие в системе ОРФО никаких помет, в современных текстах скорее могут быть результатом опечатки, но они не будут подчеркнуты.

"Подводными камнями" могут оказаться не только архаизмы.

У А. А. Блока в стихотворении "Земное сердце стынет вновь..." есть строчка "*Вернись в красивые уюты!*", которая не устраивает ОРФО тем, что "слишком много идущих подряд гласных на стыке слов". Словоформа "уюты" в современных текстах с большей вероятностью может быть результатом опечатки. Однако фраза "Комнаты очень *уюты*" с пропущенной буквой "н" у системы ОРФО не вызывает никаких комментариев.

Разговорные, просторечные, а особенно устаревшие и устаревающие слова не всегда бывают снабжены в словарях соответствующими пометами. Но желательно, чтобы при настройке, например, на режим деловой переписки такая лексика выявлялась как стилистически некорректная[8], пусть даже по субъективным представлениям разработчиков (лингвистов).

Справедливости ради следует отметить, что некоторые слова (к архаичной лексике это не относится) все же вызывают у автокорректора ОРФО сомнения с точки зрения стиля. Подсказка относит их к жаргонной, разговорной, просторечной, экспрессивной или даже бранной лексике. Однако списки таких помеченных слов в системе ОРФО далеко не полны.

В книгах, изданных полвека назад и ранее, можно встретить слова "жолтый", "чорный", "чорт" — в старой орфографии. Подобные "пережитки" встречаются и в современных текстах. Такие устаревшие варианты можно было бы включить в словарь автокорректора с соответствующими пометами.

Существуют также **просторечные и устаревшие грамматические варианты**. Например — "*с индустриализацию*": ОРФО никак не подчеркивает такие варианты творительного падежа с окончанием "-ою" или "-ею" (см. [6]). Приведем еще пример — глагольные формы на "-ся" при нормативном постфиксе "-сь" (подчеркиваются спеллером ОРФО как неопознанные). В современном русском языке такие формы носят оттенок просторечия ("а *я остаюся* с тобою" — из песни "Летят перелетные птицы" слова не выкинешь!). Об этом оттенке могла бы предупреждать подсказка автокорректора.

Встречаются и обратные случаи, которые не носят систематического характера: команды "равняйсь", "рассчитайсь". Еще примеры "сокращенных" слов: "здрасьте" вместо *здравствуйте* (в "Русском орфографическом словаре" [10] дан вариант "здрасте"), "щас" [*] вместо *сейчас*. Такие разговорные формы тоже можно было бы иметь в словаре спеллера — разумеется, с пометами.

7. "БЕЗ ГРАММАТИЧЕСКОЙ ОШИБКИ Я РУССКОЙ РЕЧИ НЕ ЛЮБЛЮ"

Эта цитата из А. С. Пушкина ("Евгений Онегин") пусть послужит нам утешением.

Грамматическими мы будем называть те ошибки, которые получаются при попытке от орфографически правильной словоформы образовать другую словоформу той же парадигмы (иногда "расширенной" парадигмы)[9]. Такие ошибки также можно было бы называть **парадигматическими**.

Иностранцы склонны образовывать множественное число и другие формы по регулярным моделям (**господины*, **гражданины*, **задают* и т. п.). Подсказка ОРФО правильных вариантов исправления (*господа, граждане, задают*) не предлагает, поскольку они отличаются от неправильных форм не одной буквой.

В устной речи и в текстах нередко встречается неправильно построенный инфинитив: **скорбить*, **бдить*, **смердить* (а также формы прошедшего времени **бдил* и т. д.) вместо *скорбеть, бдеть, смердеть* — под влиянием форм *скорбим, бдит, смердит*, **удасться* вместо *удаться* (влияние формы *удастся*), **приурАчивать* (*приурочивать*, неправильное образование парного глагола несовершенного вида от глагола *приурочить* будем тоже считать грамматической ошибкой), **проповедЫвать* (*проповедовать*).

Исключительно часто встречается пропуск мягкого знака в инфинитиве перед постфиксом "-ся". "Не хочу *учится*, а хочу *женится*" — здесь, как и в некоторых других случаях, подсказка ОРФО высказала свои сомнения. Протестируем систему еще: "Вы вправе *обратится* в суд. *Боятся* не надо. Так можно *задохнутся*. Приказано *явится завтра*". К этим фразам с пропущенным ь ОРФО не имеет замечаний[10]. Такие ошибки должны выявляться на уровне синтаксического анализа, так как все эти глагольные формы существуют.

С другой стороны, иногда "от большого усердия" появляется ненужный мягкий знак в 3-м лице настоящего и будущего времени (**надеетЬся*,

[8] Если во фразе "Давление *скакало*" пропадет буква *с* (что возможно при сканировании со сгиба развернутой книги, например), то спеллер ОРФО получившуюся глагольную форму пропустит без замечаний даже в режиме деловой переписки.

[9] Автокорректор ОРФО грамматическими называет ошибки в сочетаемости слов.

[10] Фраза с пропущенным мягким знаком "Прошу всех *удалится*" не вызывает у ОРФО возражений, а фраза "Всех прошу *удалится*" — вызывает. После недолгого тестирования удалось обнаружить, что слово "Прошу" воспринимается спеллером как имя собственное (*Проша, Проши* и т. д.).

заканчатЬся). Приведем еще пример: "Он *познакомиться* со всеми" (у спеллера ОРФО нет возражений). Эта фраза может быть получена и в результате другой ошибки — пропуска одного из предикатных слов *хочет, должен, намерен* и т. п.

Образованные нерегулярно, принадлежащие к разным моделям спряжения глагольные формы *сыпет, щипет, трепет (трепешь)* и *сыпят, щипят, трепят* уже узаконены в современных словарях в качестве вариантов к формам *сыплет, сыплют* и т. д. (см. [3, 8, 10]). Предлагавшееся написание просторечных вариантов "сыпют", "щипют" и "трепют" (см. [9]), а также (другими лингвистами) "сыпит", "щипит" и "трепит", если во множественном числе считать допустимыми формы на "-ят", не прижилось. Глагольные формы *сыпешь, сыпет* автокорректор ОРФО признает, но словоформы *сыпят, щипет, щипят, растрепет, трепется, трепят* он все же объявляет неправильными.

В романе Б. Акунина "Коронация" (изд-во Захарова, 2002 г.), опубликованном в авторской редакции, опечаток меньше, чем в среднем в книгах, изданных в последние годы. Но на стр. 8 допущена часто встречающаяся ошибка в образовании действительных причастий: **пышАщие (здоровьем)* вместо *пышущие*. Такого рода ошибки "популярны" в словах: *брызжущий, колышущийся, судачащий, внемлющий, колеблющийся, самоклеящийся* и т. п. Нередко пишут **борЯтся*, "они *надеЯтся*" при нормативных глагольных формах *борются, надеются*.

Во фразе "Это написано *чернилом*" ОРФО подчеркивает последнее слово красной волнистой линией как ошибочное, но в подсказке нормативная форма *чернилами* отсутствует. Для неправильных форм, встречающихся у не очень грамотных людей, **польта*, **хочем*, **хочете*, **хочут*, подсказка выдает "похожие" слова, отличающиеся одной буквой, и не дает правильных форм *пальто, хотим, хотите, хотят*. Некоторые ненормативные словоформы можно было бы включить с соответствующими пометами и исправлениями в список "популярных искажений".

Довольно часто можно видеть окончание "-ьи" в предложном падеже единственного числа существительных на "-ье". Нормативное написание: "Алиса в *Зазеркалье*", "на *безрыбье* и рак—рыба", "в этом *кушанье* нет соли", исключения — "*в (полу)забытьи*" как вариант (см. [3, 8, 9]). Здесь сказывается влияние существительных среднего рода на "-ие" с окончанием в предложном падеже "-ии".

Популярны формы "найм" (система ОРФО это слово "узаконила", имеет в своем словаре и не подчеркивает) и "займ" при нормативных *наём* и *заём*. Здесь мы видим влияние более употребительных косвенных падежей и "выравнивание" парадигмы. Слова "найм" и "займ" настолько часто встречаются, что уже почти превращаются в норму.

Если набрать ненормативные формы числительных **четырехстами*, **восьмистами* или **восьмидесятью*, то подсказка ОРФО предложит разные варианты исправления, но только не *четырьмястами, восьмьюстами, восемьюстами, восьмьюдесятью* (другой нормативный вариант творительного падежа *восЕмьюдесятью* [3] автокорректор подчеркивает как неправильный).

Выбор слитного или раздельного написания отрицания "не" с причастиями и прилагательными, в том числе и с краткими формами, как показывает практика, представляет большие трудности. Автокорректор ОРФО часто в таких ситуациях бывает не прав, о чем мы писали в [6]. Кстати, еще одна типичная ошибка — иногда "*не прав*" в контекстах вроде "Борис, ты *не прав*" пишут слитно.

8. ПОДНИМАЕМСЯ НА БОЛЕЕ ВЫСОКИЙ УРОВЕНЬ

Первые "русскоязычные" спеллеры, появившиеся в конце 80-х гг., проверяли только отдельные словоформы [4, 5]. Современные автокорректоры содержат уже элементы синтаксического анализа, выходят на уровень проверки словосочетаний, связей слов во фразе. В системе ОРФО есть возможность отключения любых предусмотренных проверок. Мы ссылаемся на работу спеллера при настройке на все предложенные правила.

Рассмотрим типичные **синтаксические ошибки**, ошибки в управлении глаголов, ошибки согласования и др.

Ненормативные канцеляризмы "*согласно указа*", "*согласно расписания*" мы встречаем постоянно, особенно в деловой переписке. Автокорректор ОРФО подчеркивает эти конструкции зеленой волнистой линией и указывает в подсказке, что здесь нужен дательный падеж (*указу, расписанию*).

Предлог "о" в современном языке стал употребляться где надо и где не надо. "Наше *желание о том, чтобы...*", "Это *показывает о том, что* грамотность не в чести". Система ОРФО к таким фразам не имеет значений.

Стало политкорректным говорить и писать "в Украине", а не "на Украине" (ОРФО пропускает оба варианта без комментариев).

Фразу "*Барыня приехали-с*!" ОРФО пропускает без замечаний, как, впрочем, и фразу "*Книга приехали*!".

Следующий пример не вызывает никакой реакции: "Это было *в мое день рождение*. Когда *твое день рожденье?*". Популярное выражение "*Твоя моя не понимай*" ОРФО тоже пропускает.

Склонение числительных вызывает немалые трудности, поэтому часто используется именительный (или винительный) падеж. Пример: "Начали *с тысяча пятьсот пятьдесят* рублей, потом снизили цену до ста *девяносто*". К этому предложению у спеллера ОРФО нет замечаний.

Часто вместо усилительной частицы "ни" пишут "не": "*И где бы ты НЕ был...*" (спеллер ОРФО здесь на высоте). Приведем еще примеры. "*НЕ шагу назад*!", "*А он — НЕ гугу*!", "*Не сотворим себе кумира НЕ на земле, НЕ в небесах*". Последние три фразы ОРФО пропускает без возражений.

Автокорректор ОРФО выявляет некоторые ошибки в употреблении знаков препинания. Однако сокращение с точкой при последующей прописной букве он обычно воспринимает как конец предложения и выдает сообщения о непарности скобок, если такое сокращение имеется внутри скобок, и о других несуществующих ошибках.

К сожалению, правила русской пунктуации не допускают удвоения скобки или кавычки, что затрудняет компьютерную проверку, да и просто восприятие текста. Чтобы обойти это правило, иногда используют графически разные кавычки (со скобками сложнее). Например: «*В бой идут одни*

"старики"» — название фильма о молодых летчиках.

При появлении в тексте *во-вторых* автокорректор мог бы проверять присутствие *во-первых* и т. д. Так можно обнаружить пропуск фрагмента текста. Со временем, мы надеемся, системой будут исследоваться, хотя бы в пределах абзаца, и другие **анафорические связи**.

Нереально требовать от компьютерных программ выявления любых ошибок в построении фразы. Рассмотрим пример из анекдота: "*Кто девушку ужинает, тот ее и танцует*". Во-первых, существует **переходный** глагол *ужинАть* [3, 8, 10] (от *жать—жну*) с ударением на "а́". Во-вторых, если В. Васильев танцевал мачеху в балете "Золушка", то конструкции "*ее танцует*" или даже "*тот девушку танцует*" не менее правильны в синтаксическом плане. Впрочем, если заменить в этой фразе оба глагола непереходными, имеющими в словаре [3] помету "нп", — "Кто девушку *обедает*, тот девушку *пританцовывает*", то ОРФО и в этом случае ничего не подчеркнет.

В словаре А. А. Зализняка [3] в главе "Спряжение" говорится: "...не считается проявлением переходности связь с формами В. падежа, означающими меру длительности действия или пройденное расстояние...". Можно "*обедать каждую неделю*" и "*пританцовывать всю дорогу*". На формальном уровне отличать эти конструкции с винительным падежом от конструкций "*обедать, пританцовывать девушку*" можно, исходя из того, что слово *девушка* в словаре [3] (на основе которого обычно создаются автокорректоры) имеет грамматическую помету "жо". Это означает — существительное женского рода, **одушевленное**.

В упомянутом выше романе Б. Акунина "Коронация" в главе "18 мая" (стр. 345) употреблено слово *кобчик*, которому в словарях дается толкование "птица рода соколов", вместо слова *копчик* — во фразе "Вот здесь больно, на *кобчике*". Вряд ли найдется спеллер, который сможет выловить такую лексическую ошибку, если слово *кобчик* имеется в его внутреннем словаре.

Многие неправильно построенные фразы можно выявить при разборе семантики составляющих их слов. Семантический анализ для автокорректоров — задача на ближайшее будущее.

ЗАКЛЮЧЕНИЕ

Ошибки чаще всего совершаются по аналогии с более привычными словами или в результате построения форм по более регулярным моделям. Иногда в процессе адаптации слова искажают, пытаясь сделать их "более понятными". Новые заимствования, напротив, в текстах зачастую стараются делать подчеркнуто иностранными, демонстрируя знание языков.

При разработке новых версий автокорректоров желательно учитывать статистику, а также природу происхождения ошибок и опечаток — в целях их прогнозирования, что позволит дополнять компьютерные словари списками наиболее часто встречающихся или вероятных искажений с их исправлениями.

Сервисная программа-подсказка в первую очередь должна выдавать более вероятные и адекватные варианты исправления неопознанного слова. Ранжирование, установление очередности вариантов — одна из задач для разработчиков новых версий автокорректоров. При этом желательно учитывать соответствие данному тексту (заявленному как художественное произведение, научная статья, деловая переписка и т. д.) предлагаемых подсказкой словоформ с пометами.

* * *

Автор выражает глубокую благодарность В. А. Успенскому за внимательное прочтение работы и ценные замечания.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ В СТАТЬЕ

1. Большой иллюстрированный словарь иностранных слов. — М.: "Русские словари", "Астрель", АСТ, 2002.

2. Букчина Б. З., Калакуцкая Л. П. Слитно или раздельно? Орфографический словарь-справочник. — М.: Рус. яз., 1998.

3. Зализняк А. А. Грамматический словарь русского языка. Словоизменение. — 4-е изд — М.: "Русские словари", 2003.

4. Лавошникова Э. К., Трусов А. В. Знакомьтесь, "спелл чекер" // Интеркомпьютер. — 1989. — № 2.

5. Лавошникова Э. К., Трусов А. В. От "спелл чекера" к автокорректору // Интеркомпьютер. — 1991. — № 1–2.

6. Лавошникова Э. К. О "подводных камнях" в компьютерных системах проверки правописания // Вестн. МГУ. Сер. 9. Филология. — 2002. — № 6.

7. Лавошникова Э. К. Компьютерная проверка орфографии: вчера, сегодня, завтра // Вестн. МГУ. Сер. 9. Филология. — 2003. — № 5 (в печати).

8. Орфографический словарь русского языка / Под ред. В. В. Лопатина / РАН. — М.: Рус. яз., 2000.

9. Розенталь Д. Э., Теленкова М. А. Словарь трудностей русского языка. — М.: Рус. яз., 1985.

10. Русский орфографический словарь: около 160 000 слов / Под ред. В. В. Лопатина / РАН. — М.: "Азбуковник", 2000.

Материал поступил в редакцию 18.07.03.