Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2020»

Москва, 17–20 июня 2020 г.

КОВАРНЫЕ СЛОВА И ГДЕ ОНИ ОБИТАЮТ[1]

Иомдин Б. Л. (iomdin@ruslang.ru)

Институт русского языка им. В. В. Виноградова РАН;
Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Морозов Д. А. (morozov@ruscorpora.ru)

Институт проблем передачи информации
им. А. А. Харкевича Российской академии наук, Московский
физико-технический институт (НИУ), Москва, Россия

Учебные тексты для детей призваны решать противонаправленные задачи: дети должны хорошо понимать их, но в то же время такие тексты должны учить читателей новым словам. Кажется важным иметь возможность автоматически обнаруживать слова, которые могут быть незнакомы детям разных возрастов. Сложной задачей является определение слов, которые читатели воспринимают как знакомые и понятные, но на самом деле понимают неправильно. Мы предлагаем метрику коварности слов, которая вычисляется как произведение доли тех респондентов, которые помечают слово как знакомое, на долю тех из них, которые правильно определяют его значение. Мы провели серию экспериментов и обнаружили несколько коварных слов русского языка. Мы выделили несколько гипотетических механизмов появления таких слов, отражающих близость к другим, более распространённым языковым единицам: словам, морфемам и словообразовательным моделям. Следующая задача — научиться выявлять коварные слова на основе различных языковых факторов.

Ключевые слова: корпусная лингвистика, семантика, сложность, семантическая сложность, паронимы, известность слов, членимость слова, внутренняя форма слова

DOI: 10.28995/2075-7182-2020-19-1011-1024

[1] Работа выполнена при финансовой поддержке РФФИ, проект 19-29-14224. Авторы выражают благодарность С. В. Манухиной и школьникам-участникам программы «Литературное творчество: лингвистика и русский язык» (Образовательный центр «Сириус», 2019).

DECEPTIVE WORDS AND WHERE TO FIND THEM

Iomdin B. L. (iomdin@ruslang.ru)

V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences; National Research University "Higher School of Economics", Moscow, Russia

Morozov D. A. (morozov@ruscorpora.ru)

The Institute for Information Transmission Problems (Kharkevich Institute), Moscow, Russia

Educational texts for children have two distinctly differing purposes: their readers must understand them and at the same time learn new words from them. It seems important and useful to be able to automatically detect words that may be unfamiliar to children of different ages. A challenging task is to identify words that readers perceive as familiar and understandable, but in fact understand them incorrectly. We propose a metric, called word deceptiveness, which is based on surveying and calculated as the product of the number of those respondents who mark the word as familiar by the number of those who correctly determine its meaning. We conducted a series of experiments and discovered several deceptive words in Russian. Several hypothetical mechanisms for the emergence of such words have been identified. In general, these are closeness to other, more familiar linguistic units: words, morphemes and word formation models. Future work will include an endeavor to learn to identify deceptive words on the basis of various linguistic factors.

Keywords: corpus linguistics, semantics, complexity, semantic complexity, word awareness, paronyms, inner form, word segmentability

1. Введение

> *От перца, верно, начинают всем перечить... От уксуса — куксятся, от горчицы — огорчаются, от лука — лукавят, от вина — винятся, а от сдобы — добреют. Как жалко, что никто об этом не знает...*
>
> Льюис Кэррол, Приключения Алисы в стране чудес (пер. с англ. Н. Демуровой)

Учебная литература, и в частности тексты для детей, включаемые в учебники по русскому языку и литературе, призваны решать в числе прочих две задачи, во многом противонаправленные. С одной стороны, необходимо, чтобы читатели понимали содержимое, следовательно, тексты не должны содержать

слишком сложных слов. С другой стороны, лексический состав учебников не должен быть бедным, ведь учебники — это важный источник новых слов для ученика [4]. Таким образом, важной задачей кажется поиск в учебных текстах слов, которые могут быть непонятны ученикам, в том числе с целью автоматической генерации сносок и пояснений. При этом, если ученик встречает в тексте незнакомое слово, он может и сам решить найти его значение в словаре или в другом источнике информации. Наибольший же интерес вызывают слова, которые воспринимаются носителем языка как знакомые и понятные и не вызывают желания посмотреть словарь, однако на самом деле их понимание читателем существенно отличается от понимания автора текста. В таком случае понимание конкретного эпизода или даже заметной части текста может быть искажено.

Нередко значения, воспринимаемые как «неправильные», постепенно становятся нормативными и даже вытесняют исходные. Ср. известные казусы со словами *довлеть* (первоначально ‘быть достаточным’, теперь чаще ‘нависать, тяготеть, преобладать’), *свидетель* (первоначально ‘имеющий сведения’, теперь ‘очевидец’): «Воздействие близких по звучанию слов на какое-то слово может привести (в некоторых случаях и без изменения его фонетического облика) к более или менее заметному переосмыслению его первоначального значения» [8]. Вот менее известные примеры из словаря [1]:

> Обыденный — обиходный, вседневный: *В обыденной жизни; В простом обыденном платье; Обыденный случай.* Настоящее значение однодневный, в течение одного дня сделанный, одни сутки длящийся: *Обыденная церковь в Москве и Вологде*, по преданию, построенная в одни сутки; *обыденный путь* — что можно пройти или проехать в сутки; *обыденный мотылек* — живущий одни сутки. **Примечание.** Против несообразного употребления слова *обыденный* особенно восстают знатоки языка В. Даль и Я. Грот.
>
> Каникулы — вакации (вообще время, свободное от учения). *Рождественские каникулы мы провели в деревне; На пасхальных каникулах мы вдоволь повеселились*. Настоящее значение может относиться только к жаркому летнему времени.

При этом процесс смены значения не моментален, из-за чего многие изменения фиксируются словарями с задержкой. Кроме того, большинство текстов, которые рекомендуются в качестве учебных, написаны носителями языка предыдущих поколений и часто достаточно давно. Следовательно, возникает временной разрыв между автором и читателем, что делает различие интерпретаций вполне вероятным [7].

Особый интерес представляют примеры смены значений, когда мы в сущности имеем дело с омонимами, один из которых с течением времени сменяет другой. Так, например, происходит со словом *форсить*. Это слово, образованное от заимствованного из франц. *форс*, с ударением на второй слог впервые фиксируется в словаре Даля (добавление И. А. Бодуэна де Куртенэ 1903 г.) и затем последовательно приводится словарями с толкованием ‘держаться с форсом,

важничать, выставляя что-л. напоказ; фасонить' (БТС[2]). При этом в последние годы распространяется сленговый глагол *форсить* с ударением на первый слог, заимствованный из английского языка: 'продвигать что-либо, прилагать много усилий к тому, чтобы сделать известным, популярным, постоянно предлагать для обсуждения' [2]. Мы провели эксперимент (в формате онлайн-опроса, более 1500 участников), который обнаружил прямую корреляцию между возрастом респондента и интерпретацией этого слова как форсить или форсить.

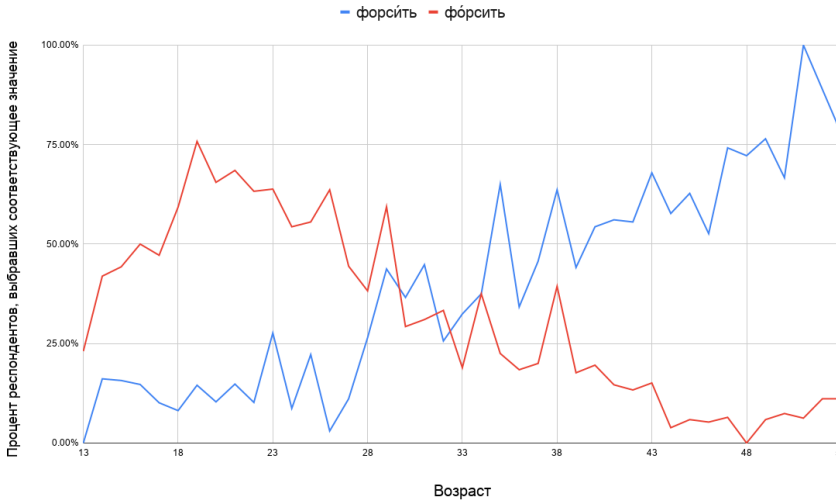

Рисунок 1. Возрастная динамика в понимании слова *форсить*

Итак, нашей целью является обнаружение ситуаций, когда слово воспринимается респондентами как знакомое, но значение, указанное ими, отличается от зафиксированного словарями. Такой случай описан, например, в работе [6] на примере слова *зябь*. Эту характеристику мы назвали *коварностью* слова. При проведении лингвистических экспериментов численное значение *коварности* для слова определялось как произведение измеренного в ходе опроса *знакомства со словом* и доли респондентов, определивших слово как знакомое и верно ответивших на вопрос о его значении. Таким образом мы оцениваем вероятность следующего события: читатель распознает слово как знакомое, но допустит при этом ошибку в понимании.

[2] БТС был выбран как один из наиболее полных толковых словарей современного русского языка; его толкования и наборы значений можно подвергнуть справедливой критике, однако для целей экспериментов, описанных в настоящей работе, это не имело существенного значния.

2. Исследование

Для поиска слов с высокой коварностью (далее «коварные слова») нами была проведена серия лингвистических экспериментов. Исходный материал был отобран в ходе масштабного опроса школьных учителей.[3] Нами были выбраны такие слова, которые неправильно понимали ученики преимущественно младших классов. Этот список был дополнен в результате совместного исследования со школьниками-участниками образовательной программы «Лингвистика» в образовательном центре «Сириус». Далее для всех этих слов были предложены гипотетические ошибочные значения. В ходе предварительного эксперимента группа школьников делала предположения о значениях слов из этого списка, а затем из указанных неверных значений для дальнейшего эксперимента выбирались наиболее частотные. Для исследования мы отбирали преимущественно слова с одним значением, поскольку полисемия существенно затруднила бы интерпретацию результатов.

Проверка знания респондентом слова проходила в два этапа. Сначала респондент указывал, насколько слово ему знакомо (Хорошо знакомо, мог(ла) бы объяснить его своими словами/Неплохо знакомо, помню контекст, в котором оно употребляется/Видел(а) когда-то, но не помню значения/Вижу это слово впервые), а затем выбирал один из четырех предложенных вариантов значения. Значение могло быть задано либо контекстом употребления, либо определением. Контексты брались преимущественно из материалов Национального корпуса русского языка (НКРЯ), а верное значение (и по возможности ошибочные, взятые из словарных статей других слов) из БТС [3].

Всего в опросах приняли участие более 750 человек, каждый из которых ответил на вопросы о 20 из исследуемых слов. Так как наибольший интерес представляют данные о коварности слов среди респондентов школьного возраста, кроме онлайн-опросов были проведены очные эксперименты в ОЦ «Сириус» (в рамках проекта лингвистической смены), в одной из школ Новосибирска и в одной из школ Москвы. Опрос прошло 259 школьников различных возрастных групп: 3–4 классы, 7–8 классы и 9–11 классы (опрос в школах), 14–15 лет и 16–17 лет (опрос в ОЦ «Сириус»). В данном исследовании мы будем обсуждать результаты респондентов именно школьного возраста. Такой выбор в том числе обусловлен тем, что опросы с их участием проходили под контролем координаторов, а это исключает влияние на результат ответов участников, пользовавшихся при прохождении дополнительными источниками информации.

---

[3] Мы использовали такой подход к подбору материала, поскольку не видим возможности более объективного подбора на данном этапе. Понятно, что читатели не смогли бы сами указать в произвольно выбранном тексте такие слова, которые понимают неконвенционально, а надежными признаками коварности слов, которые бы позволили выявить такие слова в текстах, мы пока не располагаем. Цель нашей работы и состоит в том, чтобы, изучив первоначально собранный материал, определить эти критерии и тогда уже проводить исследования на выборках из большого корпуса слов с автоматически предсказанной коварностью.

3. Результаты экспериментов

В наших экспериментальных данных были и слова, практически неизвестные респондентам (например, *деряба*[4], *пуцка, тантамареска*), и слова, воспринимаемые как хорошо знакомые (например, *роспись, символичный, замшевый, гроздь, контроллер*). При этом для многих «известных» слов респонденты давали ответы, не совпадающие с верными. Будем далее для краткости называть участников, определивших слово как знакомое или преимущественно знакомое, «опознавшими», а остальных — «не опознавшими».

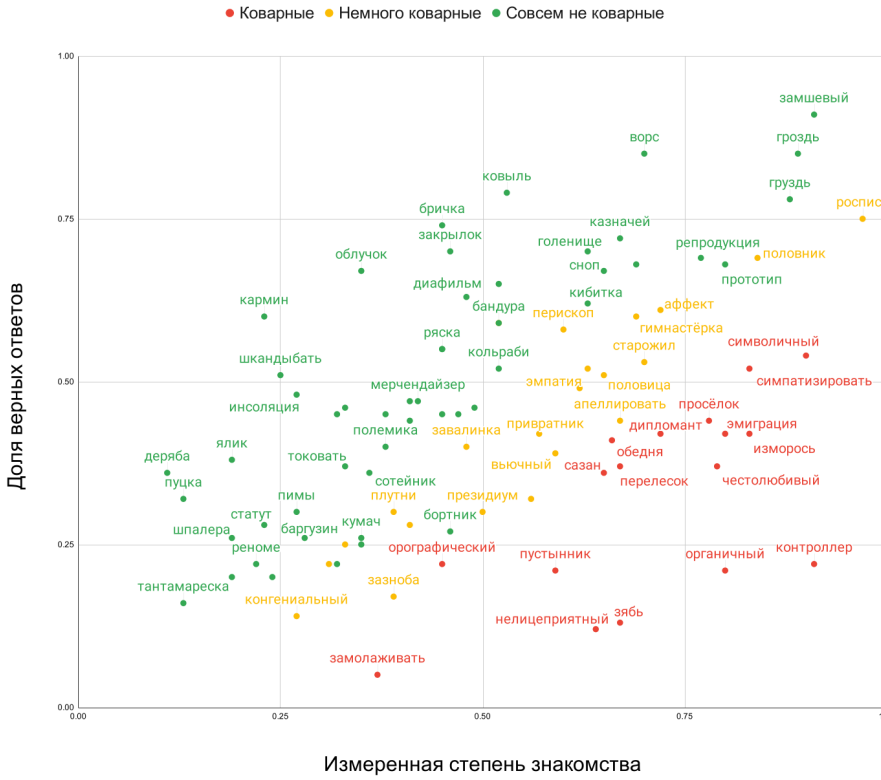

Рисунок 2. Зависимость доли верных ответов от степени заявленного знакомства со словом. Коварные слова — слова с коварностью выше 0,3, немного коварные — выше 0,2.

Мы распределили все слова из опроса, показавшие высокую коварность, на три группы по гипотетическим механизмам возникновения коварности, а затем дополнили эти группы оставшимися словами в соответствии с выделенными критериями. Рассмотрим по очереди получившиеся категории и попробуем объяснить, почему некоторые из слов в группе оказались коварными, а некоторые — нет.

[4] Слово *деряба* встречается в учебниках для начальной школы, ср. [4]

Не все из исследуемых слов вошли в какую-либо из групп. Эти нековарные слова можно разделить на две группы: незнакомые (например, *деряба*, *пуцка*) и знакомые слишком хорошо (например, *казначей*, *ворс*).

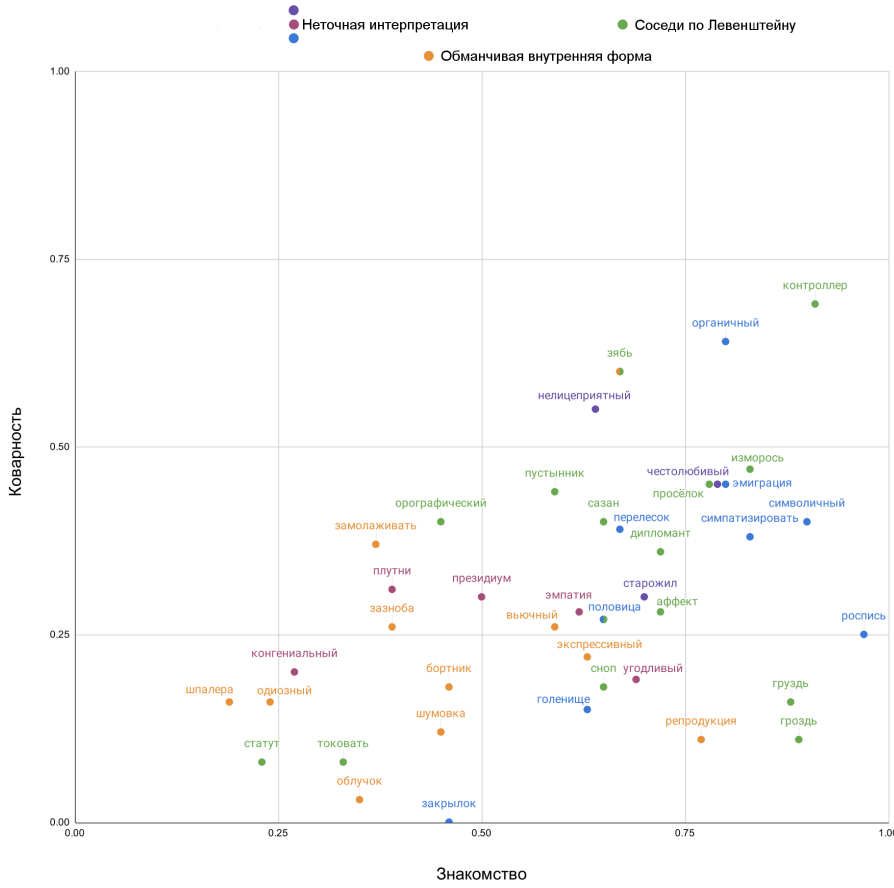

Рисунок 3. Зависимость коварности от степени заявленного знакомства со словом для различных гипотетических механизмов образования коварности

3.1. Слова, напоминающие более частотные

К этой группе мы отнесли слова, которые респонденты часто путали с более частотными словами, близкими в смысле расстояния Левенштейна [5]. При этом нередко участники опроса сами отмечали это сходство в комментариях к опросу, ср. *«Во многих словах я думал, что какая-то буква лишняя: орографический? Ф? Просёлок или посёлок? Такие слова мне кажутся сложными.»*

Ниже в **таблице 1** приведены измеренные в ходе экспериментов характеристики для слов из этой группы; слова, близкие к изучаемым в смысле расстояния Левенштейна, а также данные о частотности. Таблица упорядочена по убыванию коварности.

Следует заранее оговорить, что данные о частотности, полученные из различных корпусов, не всегда точно описывают распространённость слова в речевом опыте носителей; более того, НКРЯ в большей степени состоит из художественных и публицистических текстов, а корпус RuTenTen — из текстов, доступных в Интернете, и словарный состав этих корпусов заметно отличается от словаря школьника (собственно поэтому измеренной по корпусу частотности не всегда достаточно, чтобы предсказать, известно ли слово носителю языка школьного возраста).

Таблица 1. Слова, которые путали с близкими в смысле расстояния Левенштейна. Доля верного — доля респондентов, выбравших верный ответ (*all* — среди всех, *acq* — среди «опознавших»), доля близкого — доля «опознавших», выбравших конкретный вариант, близкий по Левенштейну, НКРЯ и RuTenTen — число вхождений для слова и близкого к нему в каждый из корпусов соответственно.

Слово	«Знакомство»	Доля верного		Коварность	Близкое слово (расстояние)	Доля близкого («опознавшие»)	НКРЯ		RuTenTen	
		all	*acq*				слово	сосед	слово	сосед
контроллер	0,91	0,22	0,24	0,69	*контролёр* (1?)	0,72	244	742	252140	72624
зябь	0,67	0,13	0,11	0,60	*рябь* (1) *зыбь* (1)	0,25	20	389 139	3704	17388 6242
изморось	0,83	0,42	0,43	0,47	*изморозь* (1)	0,43	27	104	1210	4459
просёлок	0,78	0,44	0,42	0,45	*посёлок* (1)	0,39	261	6273	9457	1277634
пустынник	0,59	0,21	0,25	0,44	*пустырник* (1)	0,29	181	42	7388	13817
сазан	0,65	0,36	0,38	0,40	*фазан* (1) *казан* (1) *каган* (2)	0,52 0,06 0,01	94	270 86 299	19603	19187 21944 16531
орографический	0,45	0,22	0,11	0,40	*орфографический* (1)	0,67	9	161	145	25203
дипломант	0,72	0,42	0,50	0,36	*дипломат* (1)	0,37	45	2473	22624	146731
аффект	0,72	0,61	0,62	0,28	*эффект* (1)	0,06	380	8760	35024	2134608
половица	0,65	0,51	0,59	0,27	*пословица* (1)	0,03	378	1084	10410	112759
актировать	0,31	0,22	0,14	0,26	*активировать* (2)	0,85	16	385	518	72969
сноп	0,65	0,65	0,73	0,18	*сноб* (1)	0,16	377	208	23762	12810
груздь	0,88	0,78	0,83	0,16	*гроздь* (1) *грусть* (2)	0,16 0,01	120	539 1708	8510	22347 107567
инсоляция	0,27	0,48	0,50	0,14	*инсталляция* (3)	0,20	29	247	8472	99206
гроздь	0,89	0,85	0,88	0,11	*груздь* (1)	0,07	539	120	22347	8510
статут	0,23	0,28	0,67	0,08	*статус* (1)	0,00	110	6987	14175	1356123

Обсудим подробнее наиболее интересные особенности полученных результатов, а также попробуем сформулировать условия для возникновения коварности слова в этой группе.

Во-первых, в большинстве коварных случаев решающую роль играло то, отделил ли респондент слово от его соседа. В этом плане показателен случай

слова *орографический*: среди «опознавших» доля тех, кто выбрал определение, соответствующее слову *орфографический*, составила 67%. В то же время, среди «не опознавших» таких респондентов всего 18%. Интересно также отметить, что относительное большинство «не опознавших» (48%) выбрало вариант значения ‘описанный в старых книгах’, по всей видимости, опираясь на присутствие *-grapho-*. Кроме того, вероятно, смешение двух слов могло отмечаться участником опроса, но тот воспринимал это как опечатку в самом тексте вопроса, иногда отмечая это в комментариях, например,: *«Наверное всё же орфографический»*.

Во-вторых, обычно «опознавшие» респонденты допускали больше всего ошибок в словах низкой частотности, причём слово-сосед должно быть частотнее.

В дополнение к этому важно, чтобы расстояние между словами было минимальным. Так, например, сопоставимы частотность и доли ошибшихся в пользу более частотного соседа «опознавших» у слов *контроллер*, *орографический* и *актировать*. Однако коварность этих слов отличается разительно, что связано с различием в уровне знакомства со словом. Мы считаем, что это связано с тем, что различие в паре *контроллер-контролёр* минимально, а удвоение или наоборот потеря удвоенной — одна из самых частых ошибок при написании в русском языке. Далее идут *орографический-орфографический* (расстояние Левенштейна 1), *актировать-активировать* (расстояние Левенштейна 2). Поэтому всё больше респондентов замечали разницу между словом и его соседом. При этом рост доли неверно ответивших «опознавших» может быть связан с тем, что среди них оказывались всё более невнимательные респонденты. Таким образом, следует отметить, что на деле нужно использовать более эффективный способ измерения воспринимаемого расстояния между словами, что может стать объектом дальнейших исследований.

3.2. Неточно интепретируемые слова

К этой группе мы отнесли слова, в которых респонденты верно выделили семантическую область слова, но всё равно допустили ошибку в понимании. В ходе обработки результатов мы решили выделить здесь три подгруппы:

1. Сложные слова с неверно определенным производящим сочетанием (*честолюбивый, старожил, нелицеприятный*).
2. Слова с непрозрачной внутренней формой (*плутни, угодливый, президиум, конгениальный, эмпатия*).
3. Слова с неверно определенной структурой семантических компонентов (*перелесок, роспись, органичный, символичный, симпатизировать, эмиграция, половица, закрылок, голенище*).

Все слова из первой подгруппы оказались достаточно известными и при этом коварными. Так, для *честолюбивый* (знакомство 0,79, коварность 0,45) 28% «опознавших» респондентов выбрало вариант ‘стремящийся сохранить честь’ вместо верного ‘стремящийся к известности, почестям’ (43%), а для *старожил* (знакомство 0,70, коварность 0,30) 23% «опознавших» респондентов

выбрали значение ‘старый человек, долгожитель’ при 56% выбравших верный вариант ‘человек, который много лет живет в одном месте’.

Прилагательное *нелицеприятный* и наречие *нелицеприятно* (знакомство 0,64, коварность 0,55) часто попадают в подборки слов с распространёнными ошибками в понимании, поэтому в этом случае мы проводили опрос как с вариантами определений, так и с вариантами толкований (каждый респондент отвечал не более чем на один из этих вопросов). В итоге в обоих вариантах опроса лидировал вариант с ярко выраженной негативной оценкой, а не словарный вариант (‘беспристрастный’):

Таблица 2. Популярность различных вариантов ответов среди «опознавших» респондентов для слова *нелицеприятный*

Вариант	Процент («опознавшие»)
Примеры употреблений	
Мой брат был лживым, нелицеприятным человеком.	61%
Он оценивал работы строго, но нелицеприятно, не выделяя своих учеников.	18%
Он был довольно нелицеприятным, но старался полюбить свою внешность.	16%
Мне было нелицеприятно выслушивать критику в свой адрес.	5%
Варианты значений	
‘крайне неприятный’	47%
‘некрасивый’	33%
‘беспристрастный’	10%
‘лживый, нечестный’	10%

Во второй группе преимущественно оказались слова с не очень высокой коварностью. Возможно, непрозрачность внутренней формы уменьшает готовность объявлять слова знакомыми. Однако присутствие в слове относительно известного корня обманывает респондента и не позволяет отметить слово как полностью неизвестное, при этом попытка восстановить смысл по казалось бы близкому слову терпит крах. Так, в случае со словом *плутни* (знакомство 0,39, коварность 0,31) «опознавшие» участники опроса в большинстве случаев выделили корень *-плут-*, однако выбирали неверный вариант, по-видимому исходя из предположения, что это множественное число слова *плут*, а не *плутня*. Аналогично со словом *президиум* (знакомство 0,50, коварность 0,30): среди «опознавших» респондентов высока доля тех, кто выбрал варианты *В новом президиуме говорится о скорой реформе образования* (38%) и *Депутат говорил речь, стоя за президиумом* (15%).

В третьей группе есть как коварные, так и нековарные слова. Одной из причин возникновения коварности в этой группе является наличие паронимичной пары или тройки: *эмиграция-иммиграция-миграция*, *органичный-органический*, *символичный-символический*, *роспись-подпись*, *половица-половик*.

Подобные случаи префиксальных и суффиксальных паронимов подробно описаны, например, в [9]. Оставшиеся коварные слова из данной группы объединены неправильным распределением участников ситуации, описываемой словом: для *симпатизировать* это произошло у 41% «опознавших» респондентов (был выбран вариант ‘нравиться кому-то’), для *перелесок* подобным образом ошиблись 42% респондентов, посчитавших слово знакомым: ими был выбран вариант *Сосновую рощу разрезали чёрные прогалины перелесков: в них не было ни единого деревца, следы пожара ещё не затянулись*, который соответствует неверному значению ‘поляны или участки без деревьев, разделяющие лес’ при словарном значении ‘небольшой лес, отделённый полянами от других лесных участков’ (последнему соответствовал вариант ответа *Долго задерживаться на лугу было опасно, и они наугад подались по перелескам на запад*, который выбрали столько же участников опроса).

Оставшиеся в этой группе слова *закрылок* и *голенище* показали низкую степень коварности. На наш взгляд, это связано с тем, что у этих слов нет ни очевидных паронимических связей, ни предпосылок для смешения участников ситуации.

3.3. Слова с обманчивой внутренней формой

К этой группе мы отнесли достаточно много слов: *ряска, ковыль, сотейник, шумовка, шпалера, замшевый, одиозный, бортник, облучок, репродукция, экспрессивный, вьючный* и другие. В соответствии с нашим предположением в подобных случаях внутренняя форма слова способна обмануть читателя: он может решить, что видит слово, однокоренное другим, хорошо ему известным, и попробует восстановить смысл, основываясь на них. Однако в реальности практически все слова этой группы либо оказались хорошо известны респондентам в верном значении, либо незнакомы. Повышенную коварность показали только слова *зябь* (0,60), *замолаживать* (0,37) и *зазноба* (0,26), причём в первом случае около половины вклада в итоговую коварность слова внёс механизм, уже упомянутый выше — близкие по Левенштейну более частотные *рябь* и *зыбь*. Все три слова крайне низкочастотные (в случае с *замолаживать* вообще практически всегда упоминается один и тот же контекст). От других слов с низкой и очень низкой частотностью в этой группе их отличает, по всей видимости, существование нескольких однокоренных слов с тем значением корня, которое респондент ошибочно приписывает слову из опроса. Так, для *замолаживать* (5 вхождений в НКРЯ) существуют слова *омолаживающий-моложавый-…*, для *зябь* (20 вхождений в НКРЯ) существуют слова *зябнуть, озябнуть, зябко, зябкий* со значением ‘мёрзнуть’, для *зазноба* (44 вхождения в НКРЯ) существуют *озноб* и *знобить*. При этом для схожего по частотности *шпалера* (104 вхождения в НКРЯ) существует фактически только одно слово — *шпала*. Аналогичная ситуация, например, со словом *сотейник* (46 вхождений в НКРЯ) — связь между существующими *соты* и *сотовый* вряд ли приходит в голову носителю.

4. Заключение

При исследовании слов, сложных для понимания читателями школьного возраста, мы обнаружили, что следует разделять случаи незнакомых и коварных слов. Если первые скорее побудят читателя каким-либо образом выяснить значение слова, то со вторыми ситуация заметно хуже: читатель скорее всего неправильно поймёт смысл конкретного слова, а вместе с ним, возможно, и какого-то отрывка текста. Поэтому коварные слова представляют больший интерес для изучения.

Для измерения коварности слова мы ввели величину *коварность* — произведение заявленного знакомства со словом и доли респондентов, определивших слово как знакомое и верно ответивших на вопрос о его значении. Такое определение позволяет нам оценить вероятность того, что слово будет некорректно распознано читателем как знакомое.

После проведения серии лингвистических экспериментов мы выявили ряд слов с высокой коварностью. В целом коварность слова, то есть его мнимая известность, обусловлена его внешней близостью к другим, более известным языковым единицам — словам, морфемам и словообразовательным моделям. Изучая полученные результаты, мы выделили несколько гипотетических механизмов возникновения коварных слов.

Во-первых, носитель может смешивать значения менее и более частотных слов, схожих по написанию и/или звучанию. В части случаев респонденты осознавали, что их ответы относятся не к данному в опросе слову, а к другому, схожему с ним, что видно по оставленным комментариям. Следует также упомянуть случай слова «контроллер», которое, по всей видимости, обладает высокой коварностью именно в письменном виде, так как в устной речи оно отличается ударением от слова «контролёр», с которым его преимущественно смешивали респонденты.

Во-вторых, слова со сложной внутренней формой могут ввести читателя в заблуждение. Здесь мы выделили две крупных группы слов: те, в которых носители выделили правильный корень, но по каким-то причинам не смогли верно интерпретировать слово целиком, и те, в которых обманчивая внутренняя форма вызвала ассоциации с корнями из другой смысловой области. Первую из перечисленных групп мы разделили на три подгруппы по типам неверной интерпретации: неправильное определение производящего сочетания для сложных слов, непрозрачная внутренняя форма и неверно понятая структура семантических компонентов.

Анализ результатов экспериментов показал, что наибольшей потенциальной коварностью обладают слова, для которых выполнено одно из следующих стечений обстоятельств:

- редкое слово с близким по Левенштейну более распространённым соседом (*контроллер*, *зябь*, *изморось*, *просёлок* и т. д.);
- относительно редкое сложное слово с распространёнными корнями и неверно понятым производящим сочетанием: *честолюбивый, нелицеприятный*;
- наличие у низкочастотного слова возможности для неправильного распределения участников ситуации, описываемой словом: *симпатизировать, перелесок*;

- существование у слова аффиксальной паронимичной пары: *символичный, эмиграция, органичный*;
- очень низкая частотность и наличие у неверно выделяемого корня способности к словообразованию: *замолаживать, зябь, зазноба*.

При этом важно помнить, что частотность слов, вычисленная по корпусу текстов, не даёт полного представления о реальной распространённости слова. Этот аспект особенно важен, так как в случае словарного запаса школьников расхождения могут оказаться значительными. Кроме того, важную роль играет контекст: слово, взятое в отдельности, может быть понято не так, как в контексте, и фонетически и/или морфологически близкие слова предположительно будут смешиваться еще сильнее, если они употребляются в схожих контекстах.

5. Дальнейшее развитие

Описанные случаи коварных слов и предполагаемые причины их коварности позволяют обозначить те области, где следует искать другие коварные слова. Однако несмотря на достигнутый прогресс, автоматический поиск случаев высокой коварности пока не представляется возможным. Для создания подобного алгоритма необходимо: во-первых, улучшить способ измерения расстояния между словами для случая слов-соседей (учитывающий кроме буквенного также фонетическое сходство слов); во-вторых, построить алгоритм, который выделял бы в слове морфемы и правдоподобные псевдоморфемы для случаев непрозрачной или обманчивой внутренней формы; в-третьих, исследовать, насколько на коварность влияет наличие контекста; наконец, в-четвёртых, необходимо изучить основные источники пополнения лексикона детей и подростков для того, чтобы улучшить оценку распространённости тех или иных слов в их речевом опыте.

Литература

1. *Dolopchev V. P.* (1909) An attempt at a dictionary of mistakes in Russian colloquial speech [Opyt slovarya nepravil'nostej v russkoj razgovornoj rechi], Typographic K. Kovalevskago, Warsaw.
2. *Krongauz M. A. et al* (2016) Internet.ru Language Dictionary [Slovar' yazyka interneta.ru] AST-PRESS KNIGA, Moscow.
3. *Kuznetsov S. A.* (1998) Great Dictionary of the Russian Language [Bol'shoy tolkovyy slovar'], Norint, St. Petersburg.
4. *Laposhina A. N., Veselovskaya T. S., Lebedeva M. U., Kupreshchenko O. F.* (2019), Lexical analysis of the russian language textbooks for primary school: corpus study [Leksicheskij sostav tekstov uchebnikov russkogo yarzyka dlya mladshej shkoly: korpusnoe issledovanie], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2019" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2019"], Moscow, Vol. 18, pp. 351–361.

5. *Levenshtein V. I.* (1965) Binary codes capable of correcting deletions, insertions and reversals, Soviet Physics Doklady, No. 8, pp. 707–710.
6. *Morozov D. A., Iomdin B. L.* (2019) Criteria of semantic complexity of words [Kriterii semanticheskoj slozhnosti slova], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2019" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2019"], Moscow, Vol. 18, Sup. vol, pp. 119–131.
7. *Sharapova E. V.* «False reader friends» in F. M. Dostoevsky «big novels» [«Lozhnye druz'ya chitatelya» v «bol'shih romanah» F. M. Dostoevskogo], Russian language and literature in professional communication and multicultural space: materials of the International scientific-practical conference [Russkij yazyk i literatura v professional'noj kommunikacii i mul'tikul'turnom prostranstve: materialy Mezhdunarodnoj nauchno-prakticheskoj konferencii], Moscow, pp. 360–364.
8. *Shmelev D. N.* (1964) Essays on the semasiology of the Russian language [Ocherki po semasiologii russkogo yazyka], Moscow, p. 184.
9. *Vischnyakova O. V.* (1984) Dictionary of Russian Paronyms [Slovar' paronimov russkogo yazyka], «Russian Language», Moscow.