КОМПЬЮТЕРНЫЙ СЛОВАРЬ РУССКИХ ПАРОНИМОВ, ОСНОВАННЫЙ НА ФОРМАЛЬНОМ КРИТЕРИИ ПАРОНИМИИ

Большакова Е. И. (eibolshakova@gmail.com)
МГУ им. М. В. Ломоносова, Москва, Россия;
НИУ Высшая школа экономики, Москва, Россия

Большаков И. А. (iabolshakov@gmail.com)
Независимый исследователь, Москва, Россия

В результате исследования наиболее крупного печатного словаря паронимов русского языка предложен формальный критерий паронимии. Паронимами считаются те пары слов одного корня и одной части речи, у которых различия в аффиксах (раздельно в префиксах и суффиксах) находятся в фиксированных рамках. Согласно этому критерию построен компьютерный словарь русских паронимов, имеющий 21,8 тыс. статей с 192 тыс. паронимов и по объему превышающий все известные словари. В первую очередь словарь предназначен для исправления в текстах ошибочных замен слов их паронимами.

Ключевые слова: паронимы, паронимия, компьютерный словарь паронимов, паронимические ошибки, исправление ошибок в текстах

A COMPUTER DICTIONARY OF RUSSIAN PARONYMS BASED ON A FORMAL CRITERION OF PARONYMY

Bolshakova E. I. (eibolshakova@gmail.com),
Lomonosov Moscow State University, Moscow, Russia;
National Research University Higher School of Economics, Moscow, Russia

Bolshakov I. A. (iabolshakov@gmail.com),
Independent researcher, Moscow, Russia

We note that Western European lexicography has neither precise definition of paronymy nor dictionaries of paronyms. However, such dictionaries can help us correct malapropisms like *massive evacuation* or sensitive shoes. Although three comprehensive dictionaries of Russian paronyms have been published in the recent decades, it remains unclear what additional features of similarity of two words of the same root and the same POS are needed to consider the words paronymous. Based on the collected statistics of affix proximity of paronyms in the largest printed dictionary of Russian paronyms, we propose a formal criterion of paronymy. Two words of the same root and the same POS are considered formally paronymous if their affix differences (separately for suffices and prefixes) are limited to particular values. Affix difference equals the minimal number of editing operations on affixes (deletion, insertion or substitution) that transform an affix chain of one word into that of the other. Aiming to develop a computer dictionary of formal paronyms, we first compiled a computer dictionary of 23,000 Russian words divided into 2,400 same-root, same-POS groups. All words were split into morphs: prefixes, the root, suffixes, and the ending. Then affix distances between word pairs from the groups were automatically computed, and all formally paronymous pairs were selected. These pairs constitute the resulting computer dictionary of paronyms, which contains 21,800 word entries with their 190,000 paronyms, larger than all known dictionaries of paronyms.

Key words: paronyms, paronymy, computer dictionary of paronyms, paronymy errors, correction of malapropisms

1. Введение

Внешнее сходство слов является источником разнообразных ошибок, встречающихся в текстах и подлежащих исправлению. В лингвистике с внешним сходством слов связаны такие понятия, как паронимы и паронимия.

В английском языке слово *paronymous* известно с середины 17 века. Но если сейчас извлечь из интернетовских сайтов десяток англо- и франкоязычных определений паронимии, то единства мнения не обнаружится. В большинстве определений указывается совпадение корня (*wise — wisdom*), в других фигурирует совпадение звучания при различии смысла и орфографии (*hare — hair*). Упоминаются также никак не уточняемые совпадение деривации, единство происхождения, различие окончаний. В словаре [7] слово *paronymous* имеет два разных смысла, и лишь один из них имеет отношение к сходству слов. Найденные определения не указывают явно принадлежность паронимов к одной части речи.

При отсутствии единства в понимании паронимии становится понятным отсутствие в западноевропейской лексикографии словарей паронимов, содержащих описание различий их значений с приведением диагностирующих контекстов (синтаксически связанных слов). Некоторые сведения о паронимах содержатся лишь в словарях и пособиях по общему словоупотреблению, например, в [5] — для английского языка.

В то же время русская лексикография за последние десятилетия дала три содержательных словаря русских паронимов [1, 6, 8]. В них много

диагностирующих контекстов, а словарь [1] указывает смысловые различия паронимов особенно детально. В предисловиях содержатся некоторые уточнения понятия паронимии. Единым является требование **одинаковости корней и частей речи.**

Однако все три словаря русских паронимов не дают строгого определения паронимии, необходимого для построения компьютерного словаря паронимов. Неясно, какие именно дополнительные черты сходства слов необходимы, чтобы они считались паронимичными. Так, [8] требует от паронимов одинаковое место ударения (например, *сытый — сытный*) и, тем самым, одинаковое число слогов, а в [1] предлагается, например, паронимическая пара *показ — показание* с весьма разным числом слогов. Остается неясным и соотношение между паронимией и семантикой сопоставляемых корней и слов в целом. Например, в [6] не признаются полноправными паронимами синонимы типа *паронимический — паронимичный* и слова с омонимичными корнями типа *платный — платяной*.

В настоящей работе понятие паронимии уточняется в связи с определенным приложением словарей паронимов. Предлагается формальный, т. е. пригодный для проверки компьютером, критерий паронимии слов русского языка, и описывается построение компьютерного словаря паронимов, отвечающих этому критерию. Исследовав наиболее полный словарь русских паронимов, созданного В. Красных [6] (далее — **К-словарь**), мы нашли ту меру сходства слов в аффиксах (раздельно в префиксах и суффиксах), которая более чем в 99 % гарантирует паронимию в ее интуитивном лингвистическом понимании. Паронимами считаются те пары слов одного корня и одной части речи, у которых различия в аффиксах находятся в фиксированных рамках.

Компьютерный словарь русских паронимов построен на базе созданного нами ранее компьютерного словаря однокоренных слов (далее — **ОКЧ-словарь**), при этом для автоматического обнаружения паронимических пар применен предложенный формальный критерий. Построенный словарь (далее — **П-словарь**) превысил по объему все известные словари русских паронимов.

2. Необходимые соглашения и уточнения

Важным приложением словарей паронимов является подбор кандидатов на исправление тех ошибок, когда в тексте одно слово заменяется на другое существующее слово, на него похожее. Такие ошибки называются малапропизмами [3, 4]. Среди малапропизмов мы рассматриваем паронимические ошибки, т. е. неправомерные замены слов на слова с тем же корнем и той же части речи, как, например, при использовании словосочетания *массивный отъезд* вместо *массовый отъезд*.

Для указанного приложения важно, чтобы паронимы отвечали **принципу морфологической инвариантности контекста**: замена в тексте одного паронима другим без внесения каких-либо иных правок не нарушает морфологическую правильность текста, хотя может изменить его смысл.

К примеру, замена *отъезд* → *поезд* в сочетании *массовый отъезд* сохраняет **морфологическую** правильность в любом контексте. В то же время замена *отъезд* → *поездка* потребовала бы пересогласования прилагательного *массовый* по роду. Поэтому сформулированный выше принцип не допускает паронимию слов *отъезд* и *поездка*.

Принцип инвариантности облегчает исправление паронимических ошибок. Так, если в тексте встретилось словосочетание *экономическая эффектность,* и каким-то способом выявлена его ошибочность, то в построенном с учетом указанного принципа словаре паронимов будут сразу найдены всевозможные замены ошибочного слова и среди них *эффектность* → *эффективность* (но не *эффект*, так как это слово другого рода). Найденной заменой текст исправляется без какого-либо дополнительного его редактирования (см. подробнее в [3, 4]).

Для соблюдения принципа инвариантности нами были приняты следующие соглашения, уточняющие понятие части речи.

Расщепление существительных по числу. Формы единственного и множественного числа одного существительного будем считать разными существительными. Возникшие пары оказываются однокоренными и попадают в ОКЧ-словаре в одну группу. Группа обычно включает четыре подгруппы: муж. рода ед. числа, жен. рода ед. числа, сред. рода ед. числа и множ. числа (для множ. числа род считается нерелевантным). Согласно принципу инвариантности, представители разных подгрупп паронимами быть не могут.

Отделение причастий от глаголов. Причастия играют в текстах синтаксическую роль прилагательных, и у них та же морфопарадигма. Поэтому мы включаем далее причастия обоих видов и залогов в прилагательные. Однословные степени сравнения прилагательных считаются отдельными прилагательными. Все это позволяют искать паронимические пары в таких группах слов, как {***стар**ый, **стар**еющий, **стар**ейший, у**стар**евающий*...}.

Отделение деепричастий от глаголов. Русские деепричастия образуют при глаголах примерно те же зависимые обстоятельственные группы, что и наречия (*уйти* → *торопясь* / *торопливо*). Поэтому мы включаем все деепричастия в наречия.

Разделение глаголов и причастий по возвратности. Стоящая за окончанием глагола или причастия (у нас — прилагательного) возвратная частица *ся/сь* существенно меняет модель управления слова, тем самым делая его непохожим на все иные слова той же части речи без частицы. Поэтому наличие / отсутствие этой частицы делит глаголы и причастия на две несопоставляемые группы.

Расщепление глаголов по виду. Два вида русского глагола могут быть существенно разными морфологически. К тому же у них есть различия в комбинаторике, например, можно *делать прыжки*, но нельзя *сделать прыжки*. Поэтому мы считаем совершенный и несовершенный виды одного глагола различными глаголами.

Следующие решения, принятые нами ещё при создании ОКЧ-словаря, связаны с пониманием одинаковости корня.

Учет алломорфизма корня. Корень слов одной группы ОКЧ-словаря может иметь несколько алломорфов (***дух*** *—* ***душ****а;* ***лиц****о —* ***лич****ина;* ***отч****ество–* ***отеч****ество*). Лишь тогда, когда алломорфы корня оказывались слишком далекими по буквенному составу, мы формировали разные группы. Например, алломорфы ***лож/лаг*** Vs. ***клад/клас*** формируют группы {*по****лож****ить, на****лож****ить, по****лаг****ать…*} Vs. {***клас****ть, вы****клад****ывать, на****клад****ывать…*}.

Включение омонимичных корней. В одну группу мы включали слова с омонимичными корнями, например: {***бур****ый,* ***бур****ный,* ***бур****овой*}. Объединяли и однокоренные слова, омонимичные корни которых имеют хотя бы один одинаковый алломорф {***душ****а,* ***дух****ота +* ***душ***}, {*за****плак****анный,* ***плач****ущий +* ***плат****ный, у****плач****енный + платяной,* ***полот****няный*}. Лишь тогда, когда объединенная группа оказывалась слишком обширной, мы разбивали ее на 2–3 группы с неперекрывающимися алломорфами корня.

Включение заимствованных слов. У заимствованных слов вычленялись иноязычные аффиксы (***ин****-/****де****-/****ре****-/****про****-дукционный;* ***ак****-кредитация*), учитывались алломорфы заимствованных корней (***дубл****-ет —* ***дупл****-ет*). Изредка русский и заимствованный корень совпадают по смыслу, и тогда в одну группу попадают, например, *ин-****нов****-ация* и *об-****нов****-ление*, вместе с их аффиксами, русскими и заимствованными.

Исключение многокоренных слов, слов с префиксоидами и суффиксоидами. Мы не рассматриваем слова с префиксоидами типа *много, едино, мульти…* и суффиксоидами типа *летн, этажн…*, а также большинство многокоренных слов. Это значит, что не считаются паронимами слова *этажный* и *многоэтажный*, *летний* и *многолетний*, *законный* и *закономерный*, а также такие слова из К-словаря, как *зловредный*, *злокачественный*, *злонамеренный*…. Однако оставлены слова, имеющие два одинаковых склеенных корня, но разные аффиксы, например: *добровольный* и *добровольческий*.

Кроме рассмотренных уточнений наше формальное определение паронимии слов учитывает их сходство в аффиксах, причем отдельно — в префиксах и в суффиксах. Аффиксное сходство двух слов будем оценивать парой целых чисел (*Np*, *Ns*). Здесь *Np* — число различающихся префиксов, т. е. минимальное количество элементарных операций редактирования цепочки префиксов (их удаление, вставка или замена), переводящих цепочку префиксов одного слова в цепочку префиксов другого слова. Аналогично определяется число *Ns* для суффиксов. Отметим, что при аффиксном сравнении слов мы считаем нерелевантными их окончания, поскольку они определяются последним словообразовательным суффиксом или корнем слова.

Ограничения на значения (*Np*, *Ns*) для паронимов установлены нами в результате статистического обследования К-словаря, в ходе которого использовались данные о морфемном разборе слов ОКЧ-словаря.

3. Морфемный разбор и морфологический анализ слов ОКЧ-словаря

ОКЧ-словарь состоит из групп слов, имеющих одинаковый корень и относящихся к одной части речи (при уточненном их понимании, описанном выше).

Поскольку омонимы имеют одинаковую морфемную структуру, омонимия слов нами не учитывается, как если бы один омоним заменяет все остальные. С учетом этого упрощения ОКЧ-словарь имеет следующий состав (январь 2013 г.):

Количество слов	**23 054**
среди них, в процентах:	
существительных	42,2
глаголов	21,9
прилагательных	33,7
наречий	2,2
Число групп	2 426
Средний объем группы	9,5
Число однокоренных пар	301 074
Число сравниваемых пар	165 719

Слова ОКЧ-словаря были подвергнуты морфемному разбору вручную (за неимением программы автоматического выделения морфов), т. е. расчленены на префиксы, корень, суффиксы и окончание. Выделять суффиксы было особенно сложно. В частности, было не ясно, как задавать окончания в инфинитивах; присоединять ли так называемые тематические гласные *а, и, е* к суффиксам причастий *ющ* и *вш*. Было понятно, что склеивание некоторых морфов, различаемых лингвистами, отнесет к паронимам множество не слишком похожих слов, а расщепление морфов сильно отдалит в пространстве аффиксов даже похожие слова. Мы не учитывали аффиксный алломорфизм, и в ряде случаев склеивали смежные суффиксы.

В число префиксов включено слитное отрицание *не*, которое наряду с префиксами *а, анти, контра, против* формирует антонимы сравниваемых слов.

Ниже представлены примеры групп ОКЧ-словаря для существительных, глаголов и прилагательных после морфемного разбора (префиксам предшествует знак «–», корню «+», суффиксу «–», окончанию «*», перед частицей *сь/ся* также ставится «–»):

```
-АК+КРЕДИТ-АЦИ*Я
-АК+КРЕДИТ-ИВ*
-АК+КРЕДИТ-ИВ*Ы
+КРЕДИТ*
+КРЕДИТ-К*А
+КРЕДИТ-К*И
+КРЕДИТ-ОВ-АНИ*Е
+КРЕДИТ-ОР*
+КРЕДИТ-ОР-К*А
+КРЕДИТ-ОР*Ы
+КРЕДИТ*Ы
```

(1a)

```
+БЕД-Н*ЕТЬ
+БЕД-ОВ*АТЬ
+БЕД-СТВ-ОВ*АТЬ
-НА+БЕД-СТВ-ОВ*АТЬ-СЯ
-О+БЕД-Н*ЕТЬ
-О+БЕД-Н*ИТЬ
-О+БЕД-Н*ЯТЬ
-О+БЕД-Н*ЯТЬ-СЯ
-ПО+БЕД-СТВ-ОВ*АТЬ
-ПРИ+БЕД-Н*ИТЬ-СЯ
-ПРИ+БЕД-Н*ЯТЬ-СЯ
```

(1b)

```
-НЕ+СИСТЕМ-АТ-ИЗ-ИР-ОВ-АНН*ЫЙ
-НЕ+СИСТЕМ-АТ-ИЧ-ЕСК*ИЙ
+СИСТЕМ-АТ-ИЗ-ИР-ОВ-АНН*ЫЙ
+СИСТЕМ-АТ-ИЗ-ИР-УЮЩ*ИЙ
+СИСТЕМ-АТ-ИЧ-ЕСК*ИЙ
+СИСТЕМ-АТ-ИЧ-Н*ЫЙ
+СИСТЕМ-Н*ЫЙ
```

(1c)

В любом слове префиксов не более трех, суффиксов — не более шести, а корень, окончание и возвратная частица единственны.

Для автоматического формирования П-словаря специальная программа дополнительно определяет необходимые морфологические категории слов в ОКЧ-словаре. Для глаголов, прилагательных и наречий находится только часть речи, а для существительных — еще число и род.

4. Статистическое обследование К-словаря

Аффиксное сходство однокоренных слов изучалось на материале К-словаря [6], который фактически служил нам обучающим массивом, воплощающим лингвистическую интуицию. В К-словаре содержится 1100 так называемых паронимических рядов из 2–7 слов. Слова паронимического ряда относятся к одной части речи (существительные, глаголы или прилагательные) и упорядочены по алфавиту. Тем самым они неявно считаются равноправными внутри своего ряда, т. е. любое из них паронимично всем остальным.

Паронимические ряды К-словаря и пары слов из одного ряда обследовались визуально. При сравнении существительных одного ряда не учитывались пары, различные по роду и/или числу, но добавлялись множественные числа тех существительных, которые таковые имеют. В глагольные ряды добавлялись глаголы другого вида, если таковой у них существует.

Для всех рассмотренных таким образом пар слов их морфемный состав брался из ОКЧ-словаря, и подсчитывались расстояния *Np* и *Ns*.

Всего в К-словаре было насчитано 3297 пар. Статистика аффиксного расстояния представлена вторым и третьим столбцами Таблицы 1. Если построить ранговое распределение статистических данных, то первые два ранга займут пары, различающиеся только одним аффиксом, причем больше всего пар различается только одним суффиксом. Третий и четвертый ранг занимают пары, различающиеся двумя аффиксами.

Неожиданно большое количество пар слов (ранг 5 статистического распределения) оказалось на минимальном расстоянии (0, 0), т. е. когда слова

имеют одинаковый морфный состав. Сюда попали существительные с алломорфизмом корня (*отечество — отчество*), глаголы и прилагательные с алломорфными корнями и/или разными окончаниями (*воскресать — воскресить — воскрешать, временный — временной*).

Легко видеть, что набор из 7 расстояний: (0, 0), (0, 1), (1, 0), (0, 2), (1, 1), (1, 2), (0, 3), выделенных в таблице контрастом, покрывает 99,5% всех рассмотренных пар; его мы и берем в качестве **критерия аффиксного сходства**. Можно записать этот критерий в виде формулы:

$$(Np = 0)\ \&\ (Ns <= 3)\ V\ (Np = 1)\ \&\ (Ns <= 2).$$

Словесная формулировка критерия такова: либо префиксы в сравниваемой паре одинаковы, а различий в суффиксах не более трех, либо у них один различный префикс, а различий в суффиксах не более двух**.** Как видим, лингвистическая интуиция составителя К-словаря допускает у паронимов больше различий в суффиксах, чем в префиксах, и предпочитает считать паронимами слова с одинаковыми началами.

Таблица 1. Статистика аффиксного сходства

	К-словарь		**ОКЧ-словарь**		
Np, *Ns*	**Число**	**%**	**Число**	**%**	**Примеры**
0, 0	144	4,3	1152	1,3	*отечество — отчество, невежа — невежда, осветить — осветлить, заспавший — засыпавший*
0, 1	1034	31,4	7267	8,0	*корона — коронка, доносить — донашивать, маленький — малый, прогулы — прогулки, двигатель — движитель*
1, 0	990	30,1	35723	39,3	*вход — выход, входить — выходить, входной — выходной, ходить — сходить, выйти — пойти*
0, 2	535	16,6	6053	6,7	*манера — манерность, актировать — активизировать, стрелковый — стреляный*
1, 1	472	14,3	23470	25,8	*аккредитация — кредитка, проведать — выведывать, означенный — назначаемый*
2, 0	4	0,1	1264	1,4	*ход — перерасход, означить — переназначить, означенный — переназначенный*
0, 3	91	2,8	848	0,9	*акт — активатор, актерствовать — активизировать, актовый — активирующий*
1, 2	10	0,3	9875	10,9	*болезнь — заболевание, актировать — дезактивировать, актированный — дезактивизированный,*

	К-словарь		ОКЧ-словарь		
Np, *Ns*	**Число**	**%**	**Число**	**%**	**Примеры**
2, 1	0	0,0	1215	1,3	*запредельность — разделенность, ходули — перерасходы, надуманный — понапридумавший*
3, 0	0	0,0	29	0,0	*деление — перераспределение, задумывать — понапридумывать*
Проч.	14	0,4	4116	4,5	*мерзость — омерзительность, политизированность — аполитичность, опубликованный — публицистический*

Таким образом, в виду крайней редкости отбрасываемых нами случаев из К-словаря, мы считаем **формальными паронимами** слова одной части речи и единого корня, аффиксное расстояние между которыми удовлетворяет указанному выше критерию.

Представленные в таблице пары, не отнесенные к формальным паронимам, как правило, брались из ОКЧ-словаря. Обычно они внешне несходны, например, пара ***ход**ули — перерас**ход**ы*. Однако слова пары ***мерз**ость — о**мерз**ительность* кажутся схожими.

5. Формирование словаря паронимов

Для автоматического построения словаря паронимов был взят ОКЧ-словарь, подвергнутый морфемному разбору и морфологической категоризации. Каждая группа словаря из *M* слов преобразуется в *M* статей: одно слово исходной группы становится головным для статьи, а *M-1* остальных, подчиненных слов упорядоченно следуют за ним. Вычисляются значения *Np* и *Ns* для всех пар <головное слово, подчиненное слово>. После отсева подчиненных слов, не отвечающих формальному критерию паронимии с головным, все статьи, в которых осталось хотя бы одно подчиненное слово, включаются в П-словарь. Например, группа (1c) однокоренных слов из семи прилагательных переходит в следующие семь статей, где число слов, паронимичных головному, колеблется от одного до четырех:

```
НЕСИСТЕМАТИЗИРОВАННЫЙ
СИСТЕМАТИЗИРОВАННЫЙ =ANT

НЕСИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧЕСКИЙ =ANT
  СИСТЕМАТИЧНЫЙ ~ANT

СИСТЕМАТИЗИРОВАННЫЙ
  НЕСИСТЕМАТИЗИРОВАННЫЙ =ANT
  СИСТЕМАТИЗИРУЮЩИЙ

СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЗИРОВАННЫЙ
  СИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧНЫЙ

СИСТЕМАТИЧЕСКИЙ
  НЕСИСТЕМАТИЧЕСКИЙ =ANT
  СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЧНЫЙ ~SYN
  СИСТЕМНЫЙ

СИСТЕМАТИЧНЫЙ
  НЕСИСТЕМАТИЧЕСКИЙ ~ANT
  СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЧЕСКИЙ ~SYN
  СИСТЕМНЫЙ

СИСТЕМНЫЙ
  СИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧНЫЙ (2)
```

Для существительных дополнительно учитывается род и число слов, и если подчиненное слово отличается от головного по этим параметрам, оно

автоматически исключается из статьи. Вот итоговые статьи, сформированные на основе группы (1а):

```
АККРЕДИТАЦИЯ        КРЕДИТ              КРЕДИТОР            КРЕДИТЫ
  КРЕДИТКА            АККРЕДИТИВ          АККРЕДИТИВ          АККРЕДИТИВЫ
  КРЕДИТОРКА          КРЕДИТОР            КРЕДИТ              КРЕДИТКИ
                                                              КРЕДИТОРЫ
АККРЕДИТИВ          КРЕДИТКА            КРЕДИТОРКА
  КРЕДИТ              АККРЕДИТАЦИЯ        АККРЕДИТАЦИЯ
  КРЕДИТОР            КРЕДИТОРКА          КРЕДИТКА

АККРЕДИТИВЫ         КРЕДИТКИ            КРЕДИТОРЫ
  КРЕДИТКИ            АККРЕДИТИВЫ         АККРЕДИТИВЫ
  КРЕДИТОРЫ           КРЕДИТОРЫ           КРЕДИТКИ
  КРЕДИТЫ             КРЕДИТЫ             КРЕДИТЫ
```

Существительное кредитование оказалось единственным в подгруппе слов среднего рода и поэтому паронимической статьи не породило. Приведенные примеры показывают, что число паронимов у разных слов из одной группы слов с одинаковым корнем и частью речи может существенно различаться.

Подсчитанная в ходе преобразования ОКЧ-словаря статистика аффиксного сходства пар слов представлена в четвертом и пятом столбцах Таблицы 1. Для сопоставимости с К-словарем были исключены наречия (их всего 2,2%). Абсолютные цифры (четвертый столбец) значительно больше, чем в К-словаре, но процентные показатели (пятый столбец) в какой-то степени схожи. Косинус второго и четвертого столбцов, рассматриваемых как векторы, равен 0,79, что дополнительно подтверждает принятый нами формальный критерий паронимии.

Результирующий словарь характеризуется следующими параметрами:

Количество статей (= головных слов)	21 802
Количество подчиненных им паронимов	192 024
Среднее число паронимов в статье	8,8

Статей стало на 7% меньше, чем слов в ОКЧ-словаре. При формировании статей произошел отсев многих однокоренных слов. Для существительных коэффициент отсева равен 3,73, для глаголов — 1,79, для прилагательных — 1,46. Тем самым, паронимия достаточно селективна, т. е. равенство корней и частей речи не гарантирует паронимию. Среднее число паронимов на статью оказалось довольно большим из-за объемных групп глаголов в ОКЧ-словаре.

Отметим, что наш критерий паронимии допускает, что паронимы могут быть синонимами (патетический — патетичный) или антонимами (типичный — атипичный), это не исключается и в К-словаре. В примере (2) показаны статьи П-словаря с автоматически размеченными (средствами компьютерного словаря [2]) синонимами и антонимами для головного слова. За исключением абсолютных синонимов, совокупности диагностирующих контекстов у таких пар слов чаще всего различны, и поэтому разумно хранить всех их в словаре паронимов. Это касается и абсолютных синонимов, имеющих разные паронимы. И только тогда, когда абсолютные синонимы образуют изолированную пару типа апельсиновый — апельсинный, их в паронимический словарь включать нецелесообразно.

6. Заключение

В результате обследования наиболее крупного печатного словаря русских паронимов мы нашли ту меру аффиксного сходства однокоренных слов, которая более чем в 99 % случаев гарантирует паронимию в ее интуитивном понимании. Предложен формальный критерий паронимии: два слова паронимичны, если имеют одинаковый корень, принадлежат одной части речи (в уточненном ее понимании), и их аффиксные различия находятся в строго установленных рамках.

С использованием формального критерия на базе компьютерного словаря однокоренных слов автоматически построен словарь русских паронимов, по объему превышающий все известные словари. Главное приложение построенного словаря — подбор слов–замен для автоматизированного исправления паронимических ошибок в текстах.

Мы не исключаем дальнейшего уточнения предложенного критерия, чтобы допустить пары с одинаковыми конечными суффиксами (мерзость — омерзительность), и в то же время исключить внешне мало похожие пары (аккредитация — кредитка). Для других приложений может потребоваться несколько иное уточнение понятия паронимии, но в любом случае исходным ресурсом для построения словаря паронимов может браться все тот же словарь однокоренных слов.

Литература

1. *Belchikov, Yu. A., Panjusheva M. S.* (2004) Dictionary of Russian Paronyms [Slovar' paronimov russkogo jazyka] Moscow, Russkij Jazyk.
2. *Bolshakov, I. A.* CrossLexica: A large electronic dictionary of collocations and semantic links between words in Russian. [KrossLeksika — bolshoj èlektronnyj slovar' sochetanij i smyslovykh svjazei russkikh slov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009" [Computational Linguistics and Intelligent Technologies: Proc. of the International Conference "Dialogue 2009"]. Moscow, 2009, pp. 45–50.
3. *Bolshakov, I. A., Gelbukh, A.* On Detection of Malapropisms by Multistage Collocation Testing. // A. Düsterhöft, B. Talheim (Eds.) Proc. 8th Intern. Conference on Applications of Natural Language to Information Systems NLDB´2003, Burg, Germany, GI-Edition, LNI V. P-29, Bonn, 2003, p. 28–41.
4. *Bolshakova, E. I., Bolshakov I. A.* (2007) Automatic detection and computer-aided correction of Russian malapropisms [Avtomaticheskoe obnaruzhenie i avtomatizirovannnoe ispravlenie russkikh malapropizmov] // Nauchnaya i Tekhnicheskaya Informatsiya. Ser. 2, No. 5, 2007, p. 8–13.
5. *Fowler, H. W.* (1994) Dictionary of Modern English Usage. Wordsworth Editions Ltd.
6. *Krasnykh, V. I.* Explanatory Dictionary of Russian Paronyms [Tolkovyj slovar' paronimov russkogo jazyka] Moscow, AST Astrel, 2007.
7. *Merriam* Webster's Collegiate Dictionary (1993) Merriam-Webster Inc.
8. *Vishnjakova, O. V.* (1984) Dictionary of Russian Paronyms [Slovar' paronimov russkogo jazyka] Moscow, Russkij Jazyk.