МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Филологический факультет

АКТУАЛЬНЫЕ ПРОБЛЕМЫ ЛИНГВИСТИКИ И ЛИТЕРАТУРОВЕДЕНИЯ

Сборник материалов VII (XXI)
Международной научно-практической конференции
молодых учёных
(16–18 апреля 2020 г.)

Выпуск 21

Томск
Издательский Дом Томского государственного университета
2020

DOI: 10.17223/978-5-94621-901-3-2020-33

**ФОРМАЛЬНЫЕ ХАРАКТЕРИСТИКИ 20 НАИБОЛЕЕ
ЧАСТОТНЫХ ОМОГРАФОВ В РАМКАХ РЕШЕНИЯ ЗАДАЧ
АВТОМАТИЧЕСКОГО СНЯТИЯ ОМОГРАФИИ
В РУССКОМ ЯЗЫКЕ**

Головина А.Л.

Томский государственный университет, магистрант

**DISAMBIGUATING THE RUSSIAN HOMOGRAPH:
FORMAL CHARACTERISTICS OF 20 THE MOST FREQUENT
HOMOGRAPHS AND THEIR USE**

Golovina A.L.

Tomsk State University, master student

В статье представлены результаты исследования 20 самых частотных омографов русского языка. Анализ выполнен на материале акцентуированного словаря и направлен на поиск характеристик, которые могут быть учтены при решении задач снятия омографии.

Ключевые слова: омографы, частотность, снятие омографии, синтез речи.

The article presents the results of classification and linguistic analysis of the 20 most frequent Russian homographs. The analysis is based on an accented dictionary and aims at ranking the characteristics that could be useful in the task of homograph disambiguation.

Key words: homographs, frequency, homograph disambiguation, speech synthesis.

Научный руководитель: К.С. Шиляев, канд. филол. наук, доцент ТГУ.

Снятие омографии является важнейшей задачей компьютерной лингвистики, а в частности – задачей синтеза речи. При некорректном прочтении слова синтезатором речи затрудняется понимание текста. Информация, полученная в ходе анализа наиболее частотных омографов, позволит более продуктивно подойти к разрешению омографии в русскоязычных текстах.

В качестве материала исследования был использован «Частотный словарь акцентуированных словоформ русского языка» [1], созданный на основе «Акцентно размеченного корпуса русского литературного языка». Из частотного словаря объемом 58749 словоформ были извлечены 2 638 омографов и соответствующие им данные о частотности. Анализ основан на двух классификациях омографов, разработанных М.Г. Петренко [2. С. 13] и Ю.Н. Гребеневой [3. С. 4]. Первый автор предлагает выделять **лексические** (*пАрить – парИть, харАктерный – характЕрный*),

лексико-грамматические (*бЕлка – белКА, бУри – бурИ*), полиграфические (*все – всѐ, смел – смѐл*), парадигматические (*гОрода – городА, сИнее – синЕе*) и стилистические (*кОмпас – компАс, рЕчушка – речУшка*) виды омографов. Второй автор разграничивает четыре группы омографов: **I группа** – слова одной части речи (*зАмок – замОк*), **II группа** – слова разных частей речи (*звОнок – звонОк*), **III группа** – разные формы одного слова (*вОлны – волнЫ*) и **IV группа** – ситуативные омографы, имя собственное и слово, также пишущееся с большой буквы, но не являющееся именем собственным (*КОли – КолИ*). Ё-омографы не учитываются в работе Ю.Н. Гребеновой и будут отнесены только к одной группе.

Анализ 20 самых частотных омографов (таблица) акцентуированного словаря показал, что, как правило, одно из противопоставляемых слов имеет значительно более низкую частотность употребления и не включается в словарь.

20 наиболее частотных омографов

№	Омограф	Частота в корпусе, ipm	2-я форма	Классификация (имя, номер группы)
1	всЁ	3418	всЕ	полиграфические
2	меня	2416	мЕня	лексико-грамматические, II
3	ужЕ	2393	Уже	лексико-грамматические, II
4	всЕ	1827	всЁ	полиграфические
5	ему	1385	Ёму	полиграфические
6	потОм	1128	пОтом	лексико-грамматические, II
7	одИн	1055	Один	лексико-грамматические, II
8	лЕт	932	лЁт	полиграфические
9	пОсле	912	послЕ	лексико-грамматические, II
10	гОда	725	годА	парадигматические, III
11	чЕм	707	чЁм	полиграфические
12	тогО	680	тОго	лексико-грамматические, II
13	пЕред	640	перЁд	полиграфические
14	годУ	472	гОду	парадигматические, III
15	тЕм	469	тЁм	полиграфические
16	нАшей	422	нашЕй	лексико-грамматические, II
17	своегО	411	своЁго	полиграфические
18	какОй	399	кАкой	лексико-грамматические, II
19	глазА	396	глАза	парадигматические, III
20	дОма	393	домА	парадигматические, III

Сюда относится вторая форма лексико-грамматических омографов III группы (*мЕня, Один, послЕ, нашЕй*) и некоторых полиграфических омографов (*лёт, ёму, Тём*). При разрешении Ё-омографии в первую очередь стоит обратить внимание на пару *все – всѐ*, так как оба слова

имеют высокую частотность употребления. Заметим, что большую часть частотных Ё-омографов составляют служебные части речи.

Более уравновешенными по частотности являются парадигматические омографы III группы: *дОма* – *домА*, *гЛАЗа* – *глазА*, *гОда* – *годА*, *гОду* – *годУ*. Не обнаружена закономерность в частотности слов, входящих в данную группу. Иначе, более частотными могут быть как существительные в единственном числе с ударением на первый слог (*дОма*), так и существительные во множественном числе с ударением на второй слог (*гЛАЗА*).

В результате анализа было выяснено, что омографы имеют две противопоставляемые словоформы. Однако они не всегда ограничиваются двумя наборами морфологических и семантических признаков. Так, например, омограф *ужЕ* с ударением на второй слог может быть выражен как существительным от слова *уж* (на длинном *ужЕ*), так и наречием (*ужЕ* *смеркалось*). Второй случай более вероятен, и противопоставляемым словом в данном случае является прилагательное в сравнительной степени *Уже* (от прилагательного *узкий*).

По результатам отнесения омографов к различным классификационным группам, мы можем сказать, что 8 слов составляют группу полиграфических омографов (Ё-омографы), 8 слов – II группу лексико-грамматических омографов (слова разных частей речи) и 4 слова – III группу парадигматических омографов (разные формы одного слова). С этой точки зрения, более полную картину покажет классификационное распределение большего числа словоформ.

Информация, полученная в ходе анализа, может быть полезна при выборе метода разрешения омографии разного вида. К омографам с редким противопоставляемым словом применимы статистические методы, тогда как в остальных случаях следует опираться на морфологические признаки или разрабатывать контекстные правила. Вместе с тем, полученные данные позволят определить, какие группы являются приоритетными в решении задач снятия омографии.

Литература

1. Венцов А.В., Грудева Е.В. Частотный словарь словоформ русского языка. Череповец : Изд-во Череповецкого гос. ун-та, 2008. URL: <http://narusco.ru/resources.htm> (дата обращения: 17.02.2020).
2. Петренко М.Г. Явление омографии в современном русском языке : автореф. дис. ... канд. филол. наук. Одесса, 1987. 16 с.
3. Гребенева Ю.Н. Словарь омографов русского языка. Ливны : Издатель Мухаметов Г.В., 2012. 278 с.