Статистический анализ и контекстуальные правила разрешения графической омонимии при синтезе речи по тексту

Statistical analysis and contextual rules of homograph disambiguation on text-to-speech synthesis

Цирульник Л. И. (L.tsirulnik@newman.bas-net.by)
Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Барбук С. Г. (sviatos@tut.by)
Минский государственный лингвистический университет, Минск, Беларусь

Лобанов Б. М. (Lobanov@newman.bas-net.by)
Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Описываются правила определения позиции ударения в омографах, основанные на результатах контекстуального и статистического анализа текстовых корпусов. Разработанные правила используются в системе русскоязычного синтеза речи по тексту «Мультифон» и позволяют повысить степень адекватности смыслового восприятия синтезированной речи.

Введение

При разработке систем синтеза речи по тексту одной из актуальных проблем является разрешение графической омонимии. Как показывает опыт создания системы русскоязычного синтеза речи по тексту, при некорректном определении позиции ударения в слове-омографе и последующем синтезе такого слова затрудняется восприятие смысла всего предложения. В качестве иллюстрации таких ситуаций можно привести отрывки из сказки П.П.Ершова «Конёк-горбунок», в которых неверно установлено ударение в омографах[1]:

«В долгом времени аль вскоре
Приключилося им горе+…»;

«Он и усом не ведёт,
На пе+чи в углу поёт,
Изо всей дурацкой мочи+:
“Распрекрасные вы очи!”»;

«Грива в землю золотая,
В мелки+ кольца+ завитая.»

«По исходе же трёх дней
Двух ро+жу тебе коней…»

Читателю предлагается самостоятельно определить, насколько сложно понять смысл такой фразы, если она произнесена с некорректным ударением в слове-омографе.

Анализу частоты встречаемости омографов в текстах русского языка и выявлению правил определения позиции ударения в омографах посвящена данная работа.

1. Классификация омографов

В основу классификации омографов, используемой для вычисления статистических характеристик, положено разбиение, предложенное в словаре [1]. Согласно этой группировке выделяются следующие классы: разные лексемы одной части речи; раз-

[1] Здесь и далее позиция ударения обозначается знаком «+» после ударного гласного

ные формы одной лексемы; разные лексемы разных частей речи; разные варианты одной лексемы.

В процессе анализа омографов было принято решение на рассматривать класс разных вариантов одной лексемы, который включает в себя следующие подклассы: один из вариантов — профессионализм, один из членов пары — допустимый вариант, один из вариантов — с пометой «в народнопоэтической речи». Очевидно, что для корректной расстановки омографов этого класса недостаточно ни статистических, ни контекстуальных правил, а требуется, в общем случае, глубокий семантико-синтаксический анализ текста.

Кроме того, из перечня омографов были исключены слова — так называемые «ё-омографы». Исследованию их статистических характеристик посвящена статья [2].

Число омографов различных категорий, используемых для проведения эксперимента, представлено в таблице 1.

Таблица 1. Количество омографов, используемых для проведения статистического анализа.

Класс омографов	Количество
Разные формы одной лексемы	1689
Разные лексемы одной части речи	1918
Разные лексемы разных частей речи	323
Общее количество омографов	3930

2. Исходные данные для проведения статистического эксперимента

Для статистического анализа были выбраны текстовые корпуса научного и художественного стилей. В качестве корпуса научного стиля текста использовались доклады конференции «Диалог-2008», содержащие 87 текстов, включающих 237543 слова.

В качестве корпуса художественного стиля текста использовались произведения современных авторов: Б. Акунина, Л. Петрушевской, Д. Рубиной. Корпус включал 8 произведений Б. Акунина общим объёмом 239 460 слов, 56 произведений Л. Петрушевской общим объёмом 87 712 слов и 5 произведений Д. Рубиной общим объёмом 52 105 слов. Общий объём корпуса составил 379277 слов.

3. Проведение статистического эксперимента

Для проведения статистического эксперимента были разработаны специальные программные средства, которые принимают на вход список омографов и текстовый корпус и позволяют вычислять следующие числовые характеристики:

1. n — количество слов в корпусе;
2. m — количество различных омографов в корпусе;
3. m_{10} — количество омографов, встретившееся в текстах более 10 раз;
4. 4. m_1 — количество омографов, встретившееся в текстах только один раз;
5. m_z — общее количество омографов, вычисляемое в соответствии с формулой

$$m_z = \sum_{i=1}^{m} h_i q_i \qquad (1)$$

где h_i — i-тый омограф,
q_i — частота встречаемости i-того омографа в корпусе.

6. p — процентное содержание омографов в текстах, вычисляемое в соответствии с формулой:

$$p_z = \frac{m_z}{n} \times 100\% \qquad (2)$$

7. d — процентное содержание в текстах омографов из входных списков, вычисляемое в соответствии с формулой:

$$d = \frac{m}{k} \times 100\% \qquad (3)$$

где k — количество омографов во входном списке.

Числовые значения k для различных классов омографов приведены выше, в таблице 1.

4. Результаты статистического эксперимента

4.1. Результаты эксперимента для общего перечня омографов

Результаты статистического анализа омографов приведены в таблице 2.

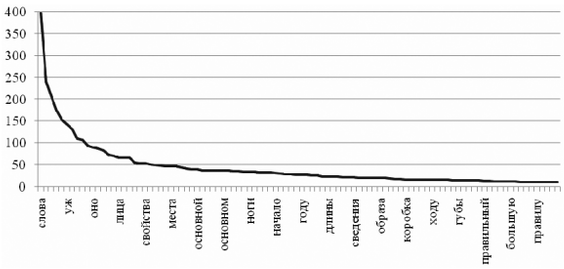
Таблица 2. Результаты статистического анализа общего перечня омографов

Стиль корпуса / Числовые характеристики	Научный	Художественный
n	237543	379277
m	537	1093
m_{10}	104	195
m_1	186	419
m_z	6089	13682
p	2,56 %	3,61 %
d	13,7 %	27,8 %

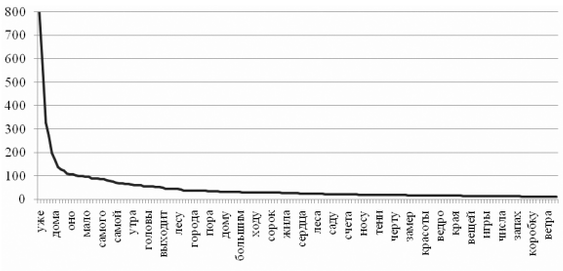
Как видно из таблицы, количество *различных* омографов, встретившихся в корпусе художественного стиля, более чем в два раза превышает количество *различных* омографов в корпусе научного стиля. В то же время около 50% омографов в корпусе художественного стиля встретилось только 1 раз[2]. В корпусе научного стиля встретилось по одному разу 34% омографов.

Интересно отметить, что, как и ожидалось, процентное содержание слов-омографов среди всех слов корпуса (величина p), а также процентное содержание омографов из входных списков (величина d) в художественных текстах выше. Такое явление можно объяснить более широким, разнообразным лексиконом текстов художественного стиля.

На рис. 1 иллюстрируется дифференциальное распределение частоты встречаемости омографов. Двадцать наиболее часто встретившихся омографов в каждом из корпусов и количество их появлений в текстах представлены в таблице 3.



а



б

Рис. 1. Дифференциальное распределение частоты встречаемости омографов:
а) в текстах научного стиля;
б) в текстах художественного стиля

Приведённые диаграммы демонстрируют более резкое «падение» частоты встречаемости омографов в корпусе художественного стиля, при этом два самых частотных омографа в художественных текстах — «уже» и «потом» — в большинстве случаев являются служебными частями речи (уже+, пото+м.).

То, что наиболее частотные омографы в корпусе научного стиля — это «слова» и «корпуса», объясняется, пожалуй, лексиконом предметной области исследуемых текстов. Шесть омографов встретились в списках двадцати наиболее частотных как в научных, так и в художественных текстах, а именно, «уже», «второй», «мало», «руки», «оно», «слова».

Таблица 3. Наиболее частотные омографы в исследуемых корпусах

Корпус научного стиля		**Корпус художественного стиля**	
Омограф	**Количество появлений в корпусе**	**Омограф**	**Количество появлений в корпусе**
слова	399	уже	811
корпуса	241	потом	555
уже	209	глаза	328
мало	176	руки	270
связи	153	голову	198
уж	143	дома	168
части	132	самом	137
правило	110	слова	127
правила	107	моя	123
стороны	95	двери	109
Оно	91	оно	108
Тона	87	деньги	108
Методы	83	должно	102
Года	74	ноги	98
Второй	71	второй	98
Лица	67	мало	97
Руки	66	окна	96
Рода	66	стороны	90
Тела	55	тому	88
Числа	54	кому	88

4.2. Результаты эксперимента для различных классов омографов

Результаты статистического анализа для разных форм одной лексемы, разных лексем одной части речи и разных лексем разных частей речи приведены, соответственно, в таблицах 4, 5, 6.

Таблица 4. Результаты статистического анализа разных форм одной лексемы

Стиль корпуса / **Числовые характеристики**	**Научный**	**Художественный**
n	237543	379277
m	155	327
m_{10}	35	73
m_1	45	100
m_z	2275	4059
p	0,96 %	1,07 %
d	9,18 %	19,36 %

[2] Эта величина вычисляется как $(m_1/m)*100\%$

Таблица 5. Результаты статистического анализа разных лексем одной части речи

Стиль корпуса / Числовые характеристики	Научный	Художественный
n	237543	379277
m	251	525
m_{10}	40	51
m_1	100	231
m_z	1746	2690
p	0,74 %	0,71 %
d	13,09 %	27,37 %

Таблица 6. Результаты статистического анализа разных лексем разных частей речи

Стиль корпуса / Числовые характеристики	Научный	Художественный
n	237543	379277
m	100	172
m_{10}	24	48
m_1	27	50
m_z	1387	3879
p	0,58 %	1,02 %
d	30,96 %	53,25 %

Как видно из таблиц, среди всех классов омографов наиболее часто (с учётом частоты появления каждого омографа) в текстах и научного, и художественного стилей встречаются разные лексемы одной части речи: 1746 и 2690 раз соответственно. При сравнении количества *различных* омографов в текстах научного и художественного стилей видно, что наибольшее их количество среди всех классов омографов принадлежит разным формам одной лексемы: 155 и 327 единиц соответственно.

Важно отметить, что процентное содержание омографов, являющихся разными формами одной лексемы и разными лексемами одной части речи, в текстах и научного, и художественного стилей практически одинаково, в то время, как процентное содержание омографов — разных лексем разных частей речи в художественных текстах в два раза больше, чем в научных.

Процентное содержание омографов из входных списков (величина p) для всех классов омографов в художественных текстах практически в два раза больше, чем в научных.

5. Контекстуальные правила

Проведённый статистический анализ позволил выявить омографы, наиболее часто встречающиеся в текстах как научного, так и художественного стилей. Затем был осуществлён экспертный анализ места ударения в таких омографах в конкретном текстовом окружении. Для этих целей использовалась специально разработанная программа, позволяющая извлекать из текстов предложения, содержащие указанные омографы, а также сайт «Национальный корпус русского литературного языка» [3], позволяющий осуществлять поиск по акцентуированному корпусу.

В результате экспертного анализа были сформулированы контекстуальные правила разрешения графической омонимии, приводимые в данном разделе.

5.1. Правила для омографов, которые являются разными формами одной лексемы

Правило 1. Для существительных женского рода третьего склонения в родительном, дательном падежах единственного числа, которые являются омографами форм слова в предложном падеже единственного числа, а также для существительных женского рода третьего склонения в именительном падеже множественного числа, являющихся омографами форм предложного падежа единственного числа, а именно: *бро+ви — брови+*, *гря+зи — грязи+*, *да+ли — дали+*, *кро+ви — крови+*, *ма+зи — мази+*, *но+чи — ночи+*, *свя+зи — связи+*, *те+ни — тени+*, *че+сти — чести+*, ударение в предложном падеже смещается с основы на окончание, если:

а. Существительное входит в несвободное словосочетание.

(1) *Примеры. Дело на мази+, всё на мази+, быть в чести+.*

б. Перед существительным находится предлог «на» (в слове «брови»).

(2) *Пример. На правой брови+.*

в. Перед существительным находится предлог «в» (в словах «дали», «ночи», «связи», «тени»).

(3) *Примеры. В дали+, в глубокой тени+.*

г. Перед существительным находится предлог «на» либо предлог «в» (в словах «грязи», «крови»).

(4) *Примеры. В грязи+, на грязи+; в тёмной крови+, на крови+.*

Исключениями из этого правила являются следующие несвободные словосочетания: «надвинуть на бро+ви», «в неразрывной свя+зи».

Правило 2. Для существительных мужского рода в форме родительного падежа единственного числа, которые являются омографами форм слова в родительном падеже единственного числа, а именно: *ря+да — ряда+, сле+да — следа+, хре+на — хрена+, ча+са — часа+, ша+га — шага+, шу+та — шута+*, ударение смещается с основы на окончание, если:

а. Это существительное второго склонения и непосредственно перед ним находятся количественные числительные «два», «три» или «четыре».

(5) *Примеры. Два ряда+, три следа+, четыре часа+, три шага+.*

б. Непосредственно перед существительным находится предлог «без» (в слове «следа»).

(6) *Пример. Без следа+.*

в. Непосредственно перед существительным или после него находится словосочетание «не осталось».

(7) *Пример. Следа+ не осталось.*

г. Существительное входит в несвободное словосочетание.

(8) *Пример. Ни хрена+.*

д. В паре омографов «шу+та — шута+» ударение падает на основу только в несвободном словосочетании «какого шу+та».

Правило 3. Для существительных мужского рода в форме дательного падежа единственного числа, родительного падежа единственного числа, которые являются омографами форм слова в предложном падеже единственного числа, а именно: *бо+ку — боку+, бы+ту — быту+, ве+тру — ветру+, ви+ду — виду+, гла+зу — глазу+, го+ду — году+, до+лгу — долгу+, до+му — дому+, ду+ху — духу+, за+ду — заду+, кра+ю — краю+, кру+гу — кругу+, ле+су — лесу+, лу+гу — лугу+, мо+згу — мозгу+, ни+зу — низу+, но+су — носу+, пле+ну — плену+, ро+ду — роду+, ря+ду — ряду+, са+ду — саду+, све+ту — свету+, сле+ду — следу+, сне+гу — снегу+, со+ку — соку+, ты+лу — тылу+, хле+ву — хлеву+, хо+ду — ходу+, цве+ту — цвету+, ча+су — часу+, ша+гу — шагу+, шка+фу — шкафу+, я+ру — яру+*, ударение в предложном падеже смещается с основы на окончание, если:

а. Перед ним находится предлог «на» либо предлог «в» (в словах «боку», «заду», «краю», «кругу», «носу», «роду», «следу», «снегу», «ходу», «часу», «шагу», «шкафу», «глазу»).

(9) *Примеры. На левом боку+, в носу+, на роду+, в глубоком снегу+, во втором часу+.*

б. Перед ним находится предлог «в» или «во» (в словах «быту», «долгу», «году», «мозгу», «плену», «ряду», «саду», «соку», «тылу», «хлеву», «цвету», «лесу», «яру»).

(10) *Примеры. В быту+, во вражьем плену+, в яру+.*

в. Перед ним находится предлог «на» (в словах «ветру», «дому», «лугу», «низу», «свету», «духу»).

(11) *Примеры. На сильном ветру+, на дому+, на духу+.*

г. Слово входит в несвободное словосочетание.

(12) *Примеры. Быть на виду+, иметь в виду+, ни в одном глазу+.*

Правило 4. Для существительных мужского и среднего рода в форме родительного падежа единственного числа, которые являются омографами форм слова в именительном падеже множественного числа (всего более 150 пар) ударение смещается на окончание, если перед ним или после него стоит глагол, прилагательное, краткое прилагательное, существительное, притяжательное или возвратное местоимение в форме множественного числа.

(13) *Примеры. Новые слова+, глаза+ смотрели.*

Правило 5. Для существительных мужского и среднего рода *антите+ла — антитела+, во+йска — войска+, дела+ — де+ла, де+ревца — деревца+, зе+ркала — зеркала+, кру+жева — кружева+, кру+жевца — кружевца+, ма+сла — масла+, ме+ста — места+, мо+ря — моря+, мы+ла — мыла+, мя+са — мяса+, о+блака — облака+, о+блачка — облачка+, по+ля — поля+, пра+ва — права+, радиозе+ркала — радиозеркала+, се+рдца — сердца+, сло+ва — слова+, ста+да — стада+, те+ла — тела+, те+льца — тельца+* ударение смещается на основу, если:

а. Перед существительным стоит количественное числительное.

(14) *Пример. Три ста+да.*

б. перед существительным стоит слово «нет» или глагол с отрицательной частицей «не».

(15) *Пример. Если для этого нет сло+ва; не было бы сло+ва; не имеете сло+ва; не знать сло+ва.*

в. перед существительным стоит местоимение в единственном числе.

(16) Пример. Этого во+йска.

Правило 6. Для существительных женского и среднего рода *ви+на — вина+*, *воло+кна — волокна+*, *гу+мна — гумна+*, *доло+та — долота+*, *ду+пла — дупла+*, *же+рла — жерла+*, *ка+йла — кайла+*, *ко+льца — кольца+*, *ли+ца — лица+*, *льноволо+кна — льноволокна+*, *о+кна — окна+*, *пи+сьма — письма+*, *поло+тна — полотна+*, *полуко+льца — полукольца+*, *полусу+кна — полусукна+*, *пя+тна — пятна+*, *ру+жья — ружья+*, *ря+дна — рядна+*, *со+пла — сопла+*, *стекловоло+кна — стекловолокна+*, *су+кна — сукна+*, *тя+бла — тябла+*, *ха+йла — хайла+*, *чи+сла — числа+*, *ядра+ — я+дра*, *яйца+ — я+йца*, *я+рма — ярма+* ударение смещается на основу, если перед существительным или после него стоит глагол, прилагательное, краткое прилагательное, существительное, притяжательное или возвратное местоимение в форме множественного числа.

(17) *Примеры. Красные ви+на, разбили о+кна, ли+ца тусклы.*

Правило 7. Для пары омографов *у+тра — утра+* ударение падает на окончание, если перед словом находится предлог «до», «от», «с» или количественное числительное.

(18) *Примеры. Девять утра+, до утра+.*

Правило 8. Для пары омографов *у+тру — утру+* ударение падает на окончание, если перед словом стоит предлог «к».

(19) *Пример. К самому утру+.*

5.2. Правила для омографов, которые являются разными лексемами одной части речи

Правило 1. Для пары омографов *сто+ит — стои+т* ударение падает на основу, если:

а. слово входит в несвободные словосочетания.

(20) *Примеры. Не сто+ит того. Этот парень коечего сто+ит!*

б. слово употреблено в безличном предложении.

(21) *Пример. Мне не сто+ит лезть в штаб-квартиру.*

Правило 2. Для пары омографов *ме+тоды — мето+ды* ударение падает на первый слог, если перед словом или после него стоит прилагательное, притяжательное или возвратное местоимение в форме множественного числа.

(22) *Пример. Математические ме+тоды.*

Правило 3. Для пары омографов *го+лову — голову+* ударение падает на окончание только в словосочетании «*городского голову+*».

5.3. Правила для омографов, которые являются разными лексемами разных частей речи

Правило 1. Для пары омографов *по+том — пото+м* ударение падает на первый слог:

а. В несвободных словосочетаниях «по+том и кровью», «умываться по+том», «умыться по+том».

(23) *Примеры. Доставалось по+том и кровью, добывать по+том и кровью.*

б. Перед или после данного слова находится глагол «покрыться», «покрываться», «облиться», «обливаться», «пахнуть», «осыпать», «осыпаться», «провонять», «залить».

(24) *Пример. После тяжёлой тренировки он обливался по+том.*

Правило 2. Для пары омографов *то+му — тому+* ударение падает на окончание, если:

а. Перед словом стоит предлог «к», а после него частица «же».

(25) *Пример. К тому+ же.*

б. Перед словом стоят предлоги «к», «по», а после них существительное в форме дательного падежа.

(26) *Примеры. К тому+ времени, к тому+ моменту.*

в. После слова стоит запятая, а после запятой одно из слов «что», «чем», «кто», «как», «какое», «с каким», «чтобы», «который», «без чего», «чему», «зачем», «о чём».

(27) *Примеры. Она не привыкла к тому+, чтобы ею командовали.*

г. Перед словом или после него находятся слова «подтверждение», «пример», «свидетель», «объяснение», «причина», «поражаясь», «поверить», «положить», «повинуясь», «накладывая», «содействовать», «верить», «готовясь к», «благодаря».

(28) *Примеры. Объяснением тому+ была плохая погода.*

Правило 3. Для пары омографов *дру+гом — друго+м* :

а. Ударение падает на основу, если слово входит в словосочетания «*друг с дру+гом*», «*друг над дру+гом*», «*друг за дру+гом*»; перед словом находится предлог «с» («со»); перед словом или после него находится местоимение или прилагательное в форме творительного падежа.

(29) Примеры. Со своим дру+гом, хорошим дру+гом был мой сосед.

б. Ударение падает на окончание, если перед словом стоят предлоги «в», «на».

(30) *Примеры. В друго+м помещении, на друго+м берегу.*

Правило 4. Для пары омографов *ми+нут — мину+т* ударение падает на окончание, если перед словом или после него находится количественное числительное, или местоимение, которое его заменяет («несколько», «столько-то», «сколько», «сколько-нибудь»), или существительное «пара».

(31) *Примеры. После двадцати мину+т, на пару мину+т.*

Правило 5. Для пары омографов *са+мом — само+м* ударение падает на основу, если слово входит в несвободные словосочетания «*в са+мом деле*», «*на са+мом деле*».

Заключение

Приведённые правила программно реализованы и применяются в системе русскоязычного синтеза речи по тексту «Мультифон» [4]. Использование описанных правил позволило повысить степень адекватного смыслового восприятия синтезированной речи.

Необходимо отметить, что описанные правила базируются только на лексико-грамматической информации о словоформах и охватывают далеко не все ситуации, встречающиеся в текстах.

Дальнейшие усилия авторов будут направлены на разработку и использование правил разрешения графической омонимии, основанных на результатах синтаксического анализа текстов.

Авторы выражают искреннею благодарность Елене Ягуновой за предоставление «Словаря омографов русского языка» [1], а также разработчикам «Национального корпуса русского литературного языка» [3].

Литература

1. *Венцов А. В., Грудева Е. В., Касевич В. Б., Корешкова Е. И., Сведенцова Е. А., Ягунова Е. В.* Словарь омографов русского языка // СПб.: Филологический факультет СПбГУ, 2004.
2. *Лобанов Б. М.* Проблема разрешения «Ё»-омографов при синтезе речи по тексту. В данном сборнике докладов.
3. *Электронный ресурс: http://www.narusco.ru/.*
4. *Лобанов, Б. М., Цирульник Л. И.* Компьютерный синтез и клонирование речи // Минск: Белорусская Наука, 2008. — 342 с.