ПРОБЛЕМА РАЗРЕШЕНИЯ «Ё»-ОМОГРАФОВ ПРИ СИНТЕЗЕ РЕЧИ ПО ТЕКСТУ

THE PROBLEM OF THE «Ё»-HOMOGRAPHS RESOLUTION IN TEXT-TO-SPEECH SYNTHESIS“

Лобанов Б.М. (*lobanov@newman.bas-net.by*),

Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь

В статье рассматривается проблема адекватного разрешения неопределенностей в системах синтеза речи по тексту, связанных с частным случаем омонимии – графической «Ё»-омонимией. Рассмотрены статистические характеристики омографических пар, в том числе «Ё»-омографов. Исследованы статистические характеристики распределений внутри наиболее часто встречающихся пар «Ё»-омографов. Обсуждаются пути разрешения наиболее частотной омографической пары «ВСЁ» и «ВСЕ».

“Когда же расставите точки над «ё»? Ё моё!!!”

LobanoPhone - 2000

Введение

Проблема адекватного разрешения неопределенностей, связанных с омонимией, играет существенную роль в решении задач распознавания и синтеза речи. Наиболее важное значение эта проблема приобретает при решении задач преобразования «речь – текст» (распознавание речи), когда существенным является разрешение почти всех видов омонимии: синтаксической, грамматической, лексической, словообразовательной и фонетической (см. словарь лингвистических терминов [1]). Только один вид омонимии - графическая омонимия, не играет роли в решении задач распознавания речи. Зато этот единственный вид омонимов, называемых омографами, играет весьма существенную роль в задачах преобразования «текст – речь» (синтез речи). Игнорирование существования омографов нарушает смысловое восприятие синтезированной речи и дополнительно ухудшает её естественность. Нам не известно ни одной работы, направленной на анализ и решение проблемы адекватного разрешения неопределенностей при синтезе русской речи по тексту, связанных с существованием омографов. В данной работе мы попытаемся в какой-то степени заполнить этот пробел, опираясь на фактический материал, представленный в словаре омографов русского языка [2].

В русском языке существуют два источника графической омонимии: вариативность *словесного ударения*, местоположение которого в письменной речи не указывается (СУ-омографы), и письменная традиция не обязательного проставления необходимых точек на *букве «Ё»* («Ё»-омографы). Литера «Ё» была предложена княгиней Екатериной Дашковой в 1783 году, а в печати употреблена в 1795 году. Отдельной буквой она долгое время не считалась и в азбуку официально не входила. В русском языке буква «Ё» используется, чаще всего в тех позициях, где произношение [(j)o] образовалось из [(j)e], чем и объясняется производная от «Е» форма буквы, хотя с точки зрения фонетики логичней было бы поставить точки не над «Е», а над «О». Букве "Ё" - 225 лет. Хотя она родилась в Санкт Петербурге, однако 20 октября 2001 года в Ульяновске

открылся единственный в мире памятник букве "Ё" (см. фото).

Существует много различных мнений, как в пользу, так и против непременного использования буквы «Ё» в печатном тексте (см. :http://www.yomaker.ru/). С нашей позиции – позиции разработчиков систем синтеза речи по тексту – отсутствие в тексте «Ё» влечёт за собой дополнительные трудности, которые должны быть разрешены в той или иной степени. Простейшее решение – игнорирование проблемы – влечёт за собой дополнительные трудности в восприятии синтезированной речи и к раздражающему слух Е-канию. Данная работа посвящена исследованию статистических закономерностей проявления «Ё»-омонимии в различных текстах, а также обсуждению вопросов разрешения связанных с ней неопределённостей.

1. Статистические характеристики омографических пар

Статистические исследования проводились с использованием специально разработанной программы “HOMOGRAPH STATISTICS” и электронного словаря омографов, созданного на основе книжного словаря [2]. Целью исследования являлось определение статистической значимости «Ё»-омографов в общем списке «СУ»- и «Ё»-омографов [2], а также выявление особенностей статистических распределений только внутри подкласса «Ё»-омографов. Общее количество омографов, в соответствии с приведенными в [2] данными, составляет 3894 пар, из них «Ё»-омографов – только 232 пары.

Статистические характеристики определялись в отдельности для достаточно представительных и различных типов текстов:

- А.С. Пушкин – стихотворные произведения,
- Л.Н. Толстой – роман «Анна Каренина,
- Б. Акунин, Д. Рубина, Л. Петрушевская – современная проза,
- Труды конференции «ДИАЛОГ-2006» - научная проза.

В таблице 1 приведены интегральные статистические характеристики этих текстов по всей совокупности омографов, содержащихся в словаре [2].

Таблица 1. Результаты теста по всем омографам

Тип текста	**Общее количество слов в тексте**	**Общее количество пар омографов**	**Число различных пар омографов**
Словарь омографов [2]	**-**	***3894 (100%)***	***3894 (100%)***
А.С. Пушкин	266.726 (100%)	9.421 (3,53%)	827 (21,2 %)
Л.Н. Толстой	279.448 (100%)	8.747 (3,13%)	680 (17,5%)
Б. Акунин и др.	379.277 (100%)	13.630 (3,59%)	1088 (27,9%)
«ДИАЛОГ-2006»	305.742 (100%)	7.195 (2,35%)	563 (14,5%)
Среднее количество	**307.775 (100%)**	**3,15%**	**20,3%**

Как видно из таблицы 1, выбранные тексты различных жанров имеют примерно одинаковый объём, в среднем – около 300 тыс. слов. Средний процент вхождения омографов составил 3,15%. Если считать, что среднее число слов на странице равно 650, то около 20-ти слов могут оказаться омографами. В случае их неадекватного раскрытия, как показывает опыт, это приводит к весьма негативному впечатлению при прослушивании синтезированной речи. Из таблицы видно также, что наибольшее количество омографов встречается в современной прозе, а наименьшее – в научном тексте. Очень интересный факт вытекает при рассмотрении 4-

го столбца таблицы: всего только порядка 20% от общего многообразия всех омографических пар встречается в проанализированных текстах! Это указывает на первостепенную важность этого подмножества в решении задач разрешения омографии.

В таблице 2 приведены статистические характеристики 4-х классов текстов по совокупности пар «Ё»-омографов, содержащихся в словаре [2].

Таблица 2. Результаты теста по «Ё»-омографам

Тип текста	**Общее количество слов в тексте**	**Общее количество пар «Ё»-омографов**	**Число различных пар «Ё»-омографов**
Словарь омографов [2]	**-**	***232 (100%)***	***232 (100%)***
А.С. Пушкин	266.726 (100%)	1.411 (0,53%)	71 (30,6%)
Л.Н. Толстой	279.448 (100%)	2.276 (0,81%)	56 (24,1%)
Б. Акунин и др.	379.277 (100%)	2.935 (0,77%)	82 (35,3%)
«ДИАЛОГ-2006»	305.742 (100%)	810 (0,26%)	49 (21,1%)
Среднее количество	**307.77 (100%)**	**0,59%**	**27,8%**

В сравнении с данными таблицы 1, средний процент вхождения «Ё»-омографов значительно ниже и составил 0,59%, что соответствует их общему количеству. Однако, если сравнить отношение количества всех пар омографов к количеству «Ё»-омографов: ***3894/232*=16,8** и соответствующее отношение процентов их вхождения в тексты: **3,15/ 0,59=5,3**, то можно отметить более чем 5-ти кратную частотность «Ё»-омографов, а, следовательно, существенную важность разрешения этого вида омографии при синтезе речи. Как и в случае таблицы 1, только порядка 30% от общего многообразия всех «Ё»-омографических пар встречается в проанализированных текстах.

В таблице 3 приведены дифференциальные характеристики статистического анализа текстов по всей совокупности омографов (первые 15 наиболее частотных пар омографов), содержащихся в словаре [2]. Как видно из таблицы, во всех художественных текстах пара «Ё»-омографов слова «***все***» выдвинулась на 1-е место. В специфическом научном тексте «Диалог-06» омограф «***все***» уступил 1-е место, к нашему удовольствию, омографу «***слова***». Из таблицы видно также, что и некоторые другие «Ё»-омографы вошли в число наиболее частотных: **«*перед, всем*».** На рисунке 1 графически представлены распределения количества встречаемости в различных текстах 10-ти наиболее частотных пар омографов. Из рис. 1 видно, что пары омографов наиболее равномерно распределены (*а, следовательно, наиболее информативны!*) в стихотворных произведениях А.С. Пушкина и в научных трудах участников «ДИАЛОГа».

Таблица 3. Результаты теста по всем омографам

А.С. Пушкин		**Л.Н. Толстой**		**Б. Акунин**		**Диалог-06**	
все	458	***все***	1670	***все***	1963	*слова*	735
уж	436	*уже*	601	*уже*	811	***все***	433
уже	361	*надо*	376	*потом*	555	*уже*	247
перед	214	*потом*	229	*уж*	345	*связи*	184
моя	204	*глаза*	211	*глаза*	328	*части*	133
всем	132	*слова*	173	*надо*	319	*корпуса*	133
глаза	126	*уж*	164	*руки*	270	*стороны*	125
сердца	120	*тому*	144	***перед***	265	*правила*	124
слова	113	*голову*	143	*голову*	198	*правило*	118

потом	112	*руки*	143	*дома*	168	*оно*	114
ночи	108	***всем***	125	***всем***	141	***перед***	105
тому	98	*дома*	124	*самом*	137	*тона*	103
пора	95	*лица*	112	*слова*	127	*рода*	101
души	95	*дела*	100	*моя*	123	*второй*	93
мою	92	*должно*	91	*двери*	109	*свойства*	91

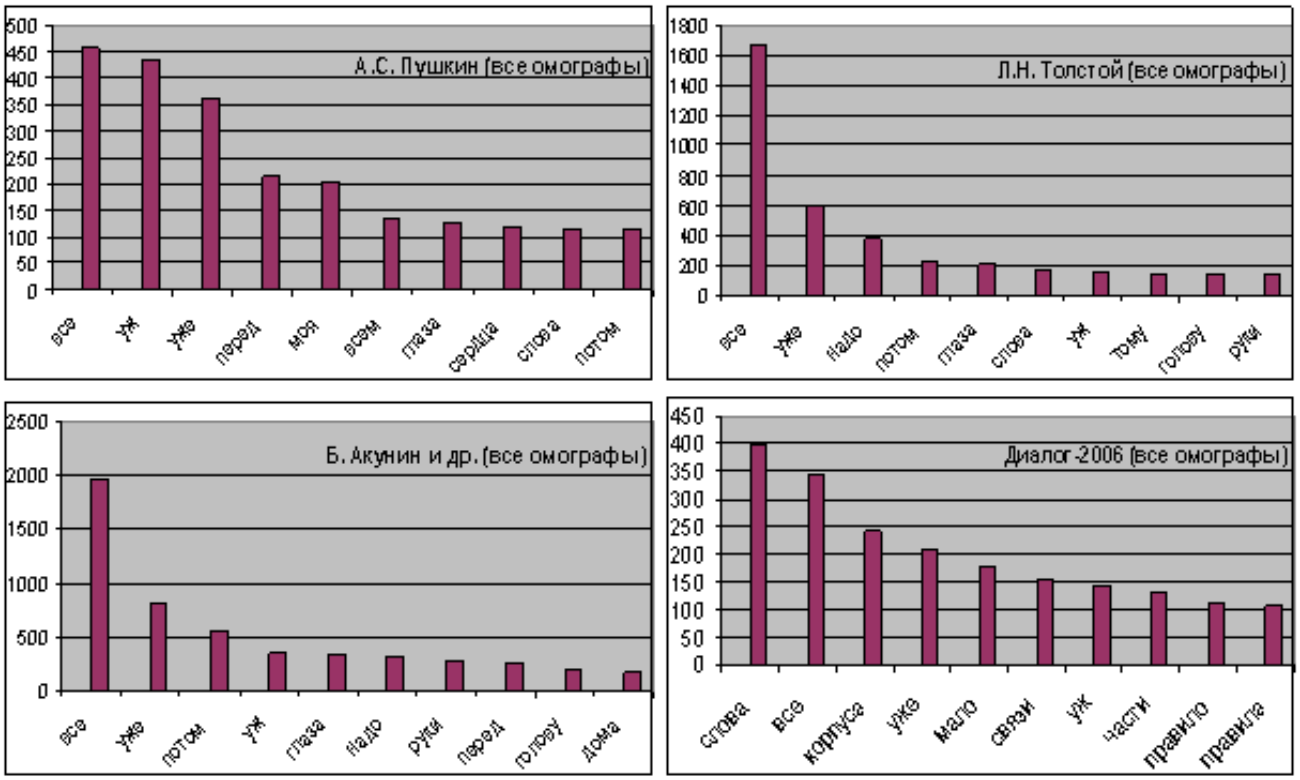

Рис.1. Распределения встречаемости пар омографов в различных текстах

В таблице 4 приведены дифференциальные статистические характеристики текстов – первые 15 наиболее частотных пар «Ё»-омографов, содержащихся в словаре [2]. Как и ожидалось 1-е места во всех текстах заняла пара омографов ***«все»***. Соответствующие таблице графичекие распределения представлены на рис. 2.

Таблица 4. Результаты теста по «Ё»-омографам

А.С. Пушкин		**Л.Н. Толстой**		**Б. Акунин**		**Диалог-06**	
все	458	***все***	1670	***все***	1963	***все***	433
перед	214	***всем***	125	***перед***	265	***перед***	105
всем	132	***жены***	78	***всем***	141	***всем***	41
слезы	92	***слезы***	53	***слезы***	56	*объем*	39
небо	50	***села***	45	***чем-то***	37	*падеж*	36
села	42	***перед***	35	***небо***	34	*пометы*	20
жены	37	***чем-то***	32	***села***	32	***небо***	14
берег	29	***сестры***	27	*щеки*	28	*помет*	12
лета	28	*умел*	22	***жены***	26	***берег***	10
весны	23	*чем-нибудь*	17	***сестры***	23	***чем-то***	9
умел	22	***небо***	14	*счета*	21	*ребра*	7
небо	21	*щеки*	11	*умел*	17	*запрет*	7
смел	18	*звезды*	9	*стекла*	17	***жены***	6

небом	15	*весел*	9	*небе*	14	*полет*	6
небе	14	*весны*	8	*осел*	14	*черта*	6
берет	13	*черта*	8	*черта*	13	*села*	4

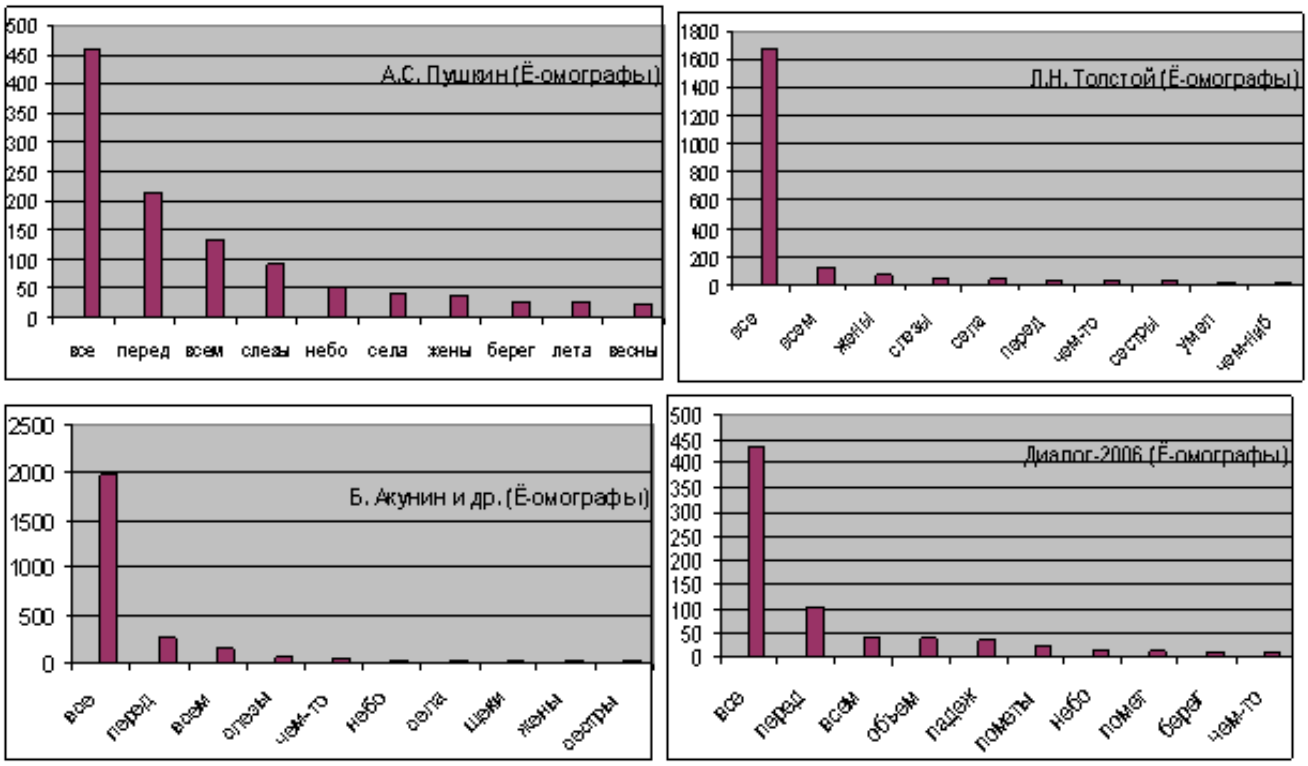

Рис.2. Распределения встречаемости пар «Ё»- омографов в различных текстах

2. Статистические характеристики распределений внутри пар «Ё»-омографов

Для определения статистических характеристик распределений внутри пар «Ё»-омографов использовались результаты описанного выше статистического анализа дифференциальных характеристик пар «Ё»-омографов и данные Интернет ресурса [3] "*Поиск по акцентуированному корпусу*". Вначале из таблицы 4 были отобраны 10 наиболее частотных пар «Ё»-омографов по всем рассмотренным выше ***4-м текстам*** (помечены жирным шрифтом в табл. 4) и подсчитаны суммарные количества их встречаемости (см. столбец 2 таблицы 5 и рис.3). Затем для этих слов с помощью Интернет ресурса [3] в ***Корпусе*** текстов по драматургии, беллетристике, публицистике и научно-популярной литературе определены суммарные количества их встречаемости (см. столбец 3 таблицы 5 и рис. 3). В столбцах 4, 5 приведены результаты встречаемости в Корпусе [3] «Ё» и «Е» слов (см. также рис. 4), в столбцах 6, 7 – соотношение количества слов с «Ё» и «Е» в процентах внутри пар «Ё»-омографов (см. также рис. 5).

Таблица 5. Парная и внутрипарная встречаемость «Ё»-омографов

Пара «Ё»-омографов	**Кол. пар в 4-х текстах**	**Кол. пар в Корпусе**	**Кол. Ё-слов в Корпусе**	**Кол. Е-слов в Корпусе**	**Соотношение внутри пар**	
					% кол. «Ё»	**% кол.«Е»**
1	***2***	***3***	***4***	***5***	***6***	***7***
все	4524	5970	4143	1826	**100**	**44**
перед	620	640	0	640	0	100
всем	440	505	109	362	**28**	**100**
слезы	200	60	60	1	100	2

села	120	64	2	62	3	100
небо	100	126	0	126	0	100
чем-то	80	123	53	70	**75,7**	**100**
жены	64	49	14	35	**40**	**100**
сестры	52	34	24	10	**100**	**42**
берег	40	85	4	81	5	100

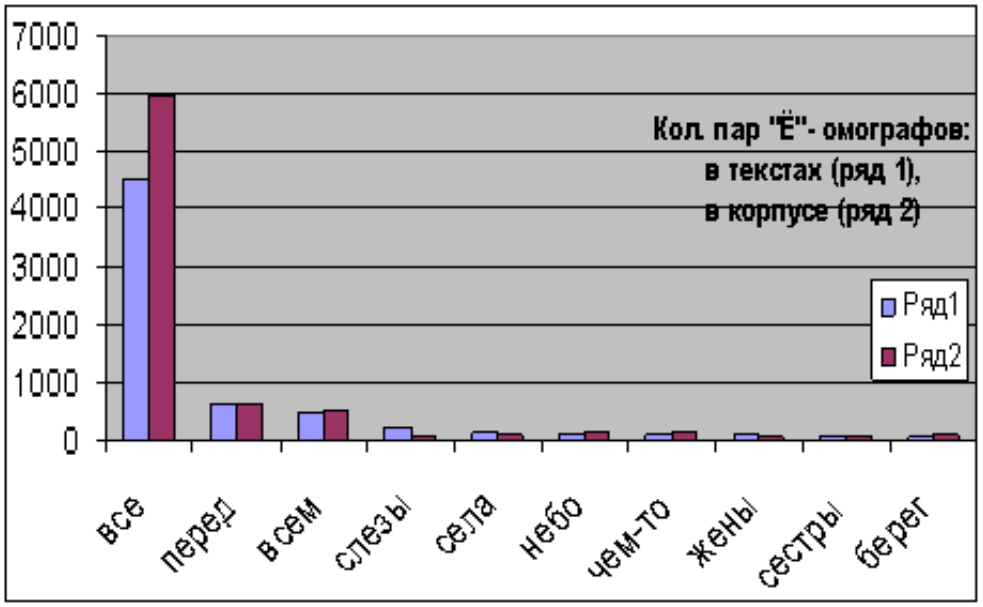

Рис.3. Распределения встречаемости 10-ти наиболее частотных пар «Ё»-омографов

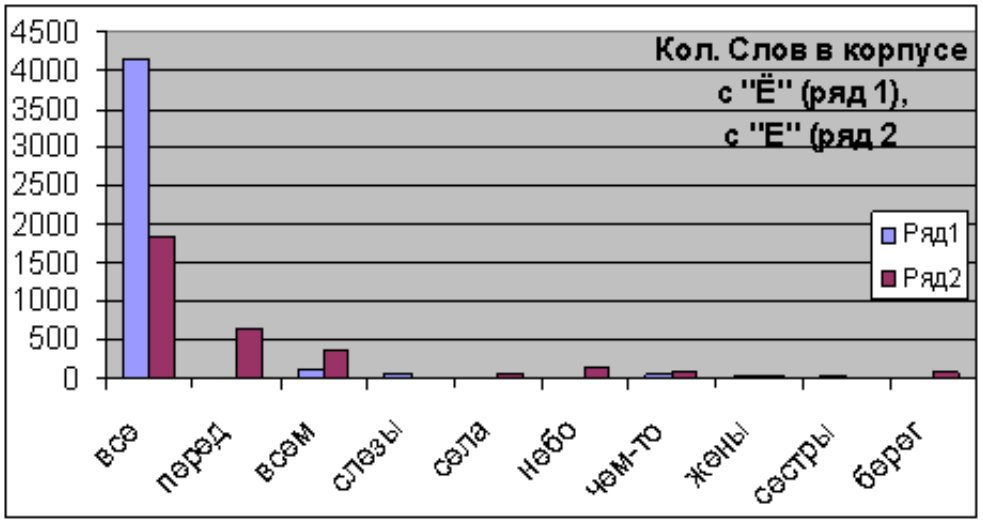

Рис. 4. Распределения кол. слов с «Ё» (ряд 1)- и «Е»(Ряд 2) внутри пар «Ё»-омографов

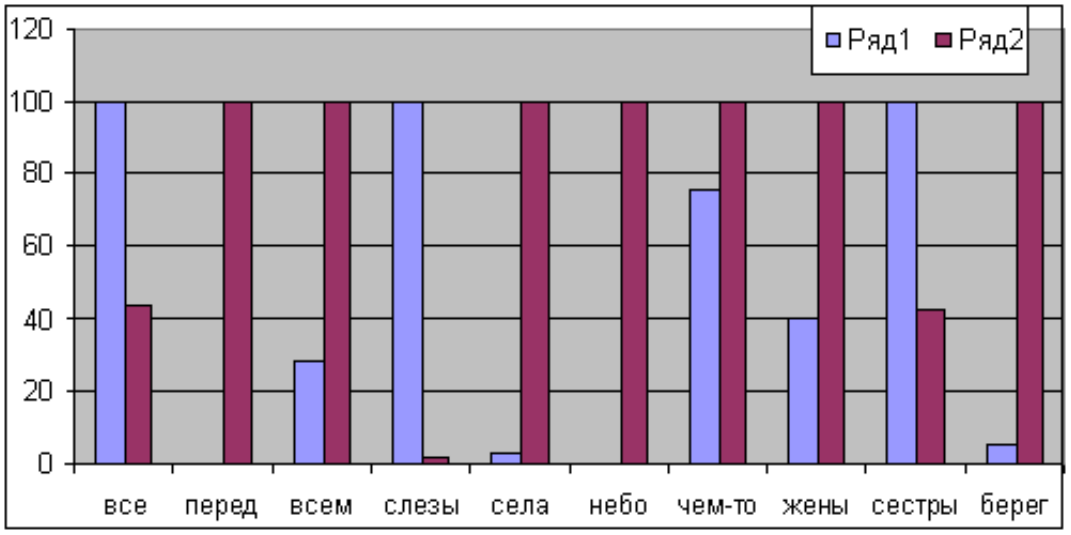

Рис. 5. Соотношения кол. слов в % с «Ё» (Ряд 1) и «Е» (Ряд 2) внутри пар «Ё»-омографов

3. Некоторые правила разрешения «Ё»- омографической неопределённости

Анализируя результаты, приведенные в таблице 5 и на рис. 3 – 4, можно сделать следующие выводы.

1. Как видно из табл. 5 (столбцы 2 и 3) использованная для статистического анализа выборка ***Текстов*** (А.С. Пушкин – стихотворные произведения, Л.Н. Толстой – роман «Анна Каренина, Борис Акунин, Дина Рубина, Людмила Петрушевская – современная проза, Труды конференции «ДИАЛОГ-2006» - научная проза) является достаточно представительной и сравнимой по объёму с ***Корпусом*** текстов по драматургии, беллетристике, публицистике и научно-популярной литературе, представленном в [3].
2. Полученные распределения встречаемости 10-ти наиболее частотных пар «Ё»-омографов в изученных ***Текстах*** и в ***Корпусе*** в высокой степени подобны (см. рис. 3), что говорит о достаточной степени достоверности полученных данных.
3. Из рис. 3 и 4 следует, что подавляющее количество «Ё»-омографов как ***Текстах,*** так и в ***Корпусе*** приходится на пару омографов «ВСЕ», что подчёркивает исключительную важность нахождения правил их разрешения при синтезе речи.
4. Из табл. 5 (столбцы 6, 7), а также из рис. 5 видно, что в 5-ти из 10-ти наиболее частотных пар «Ё»-омографов появление той или иной реализации омографа в паре более или менее равновероятно (пары: *ВСЁ_ВСЕ, ВСЁМ_ВСЕМ, ЧЁМ-ТО_ЧЕМ-ТО, ЖЁНЫ_ЖЕНЫ, СЁСТРЫ_СЕСТРЫ*). В оставшихся 5-ти парах с высокой степенью достоверностью можно выбирать варианты: *ПЕРЕД, СЛЁЗЫ, СЕЛА, НЕБО, БЕРЕГ*.
5. Для пар омографов: *ВСЁМ_ВСЕМ, ЧЁМ-ТО_ЧЕМ-ТО,* слова с «Ё» с высокой степенью достоверностью могут быть определены по наличию перед ними предлогов *«о», «об»* или *«обо»*.
6. Для пар омографов: *ЖЁНЫ_ЖЕНЫ, СЁСТРЫ_СЕСТРЫ,* слова с «Ё» могут быть определены по их пренадлежности к существительным множественного числа.
7. Наибольшую трудность представляет разрешение омографической неопределённости для слов *ВСЁ_ВСЕ*.

3.1.1. «ВСЁ» или «ВСЕ»?

Для разрешения омографической неопределённости пары ***ВСЁ_ВСЕ*** можно использовать некоторые эмпирически найденные контекстуальные правила, работающие с достаточно высокой степенью достоверностью. Для этой цели был проведен выборочный анализ встречаемости слов ***ВСЁ*** и ***ВСЕ*** в сочетании с другими словами в романе Б. Акунина «Азазель», содержащего 55 тыс. слов. Было подсчитаны количество сочетаний слова *ВСЁ* с различными словами или знаками препинания при условии, что слово *ВСЕ* ни разу не встретилось в тех же сочетаниях. Получены следующие наиболее частотные сочетания этого вида:

- *ВСЁ+Любой Знак Препинания* – 24 раза
- *ВСЁ+РАВНО* – 21раз
- *ВСЁ+ ЭТО* – 11 раз
- *ВСЁ+ТАК(ТОТ, ТЕМ) ЖЕ* – 9 раз
- *ВСЁ ВРЕМЯ* – 5 раз
- *ВСЁ ЕЩЁ* – 4 раза
- *ВСЁ БЫЛО* – 3 раза
- *ВСЁ МОЖЕТ* – 3 раза.

Определено также около 30 других сочетаний такого рода, встретившихся от 1-го до 2-х раз в проанализированном тексте.

Для более глубокого анализа возможностей разрешения омографической неопределённости пары ***ВСЁ_ВСЕ*** на том же тексте были проведены эксперименты с использованием синтаксического разбора предложений с использованием разработанной в Институте проблем передачи информации РАН системы ЭТАП-3, которая для каждого предложения строит синтаксическую струкутуру в виде дерева зависимостей [4]. На рис. 6 – 8 приведены примеры правильного синтаксического разбора предложения со словом *ВСЁ*. При правильном разборе омограф ***ВСЁ*** маркируется либо как местоимение-существительное (S) единственного числа среднего рода (рис.6), либо как местоимение-прилагательное (A) единственного числа среднего рода (рис. 7), либо как частица (PART), играющая роль ограничителя (рис. 8).

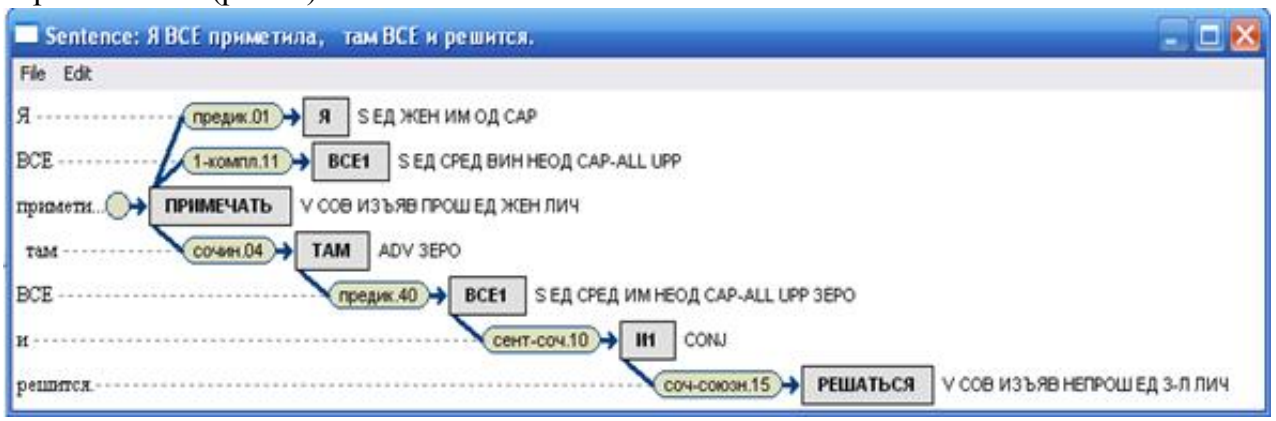

Рис. 6. Пример 1 правильного синтаксического разбора предложения со словом ***ВСЁ***

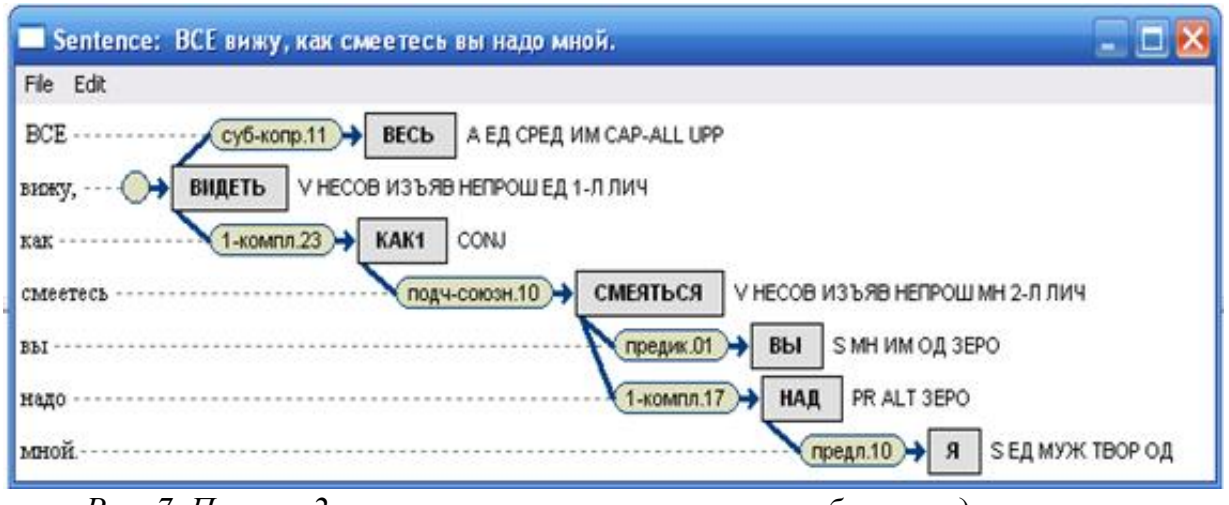

Рис. 7. Пример 2 правильного синтаксического разбора предложения со словом ***ВСЁ***

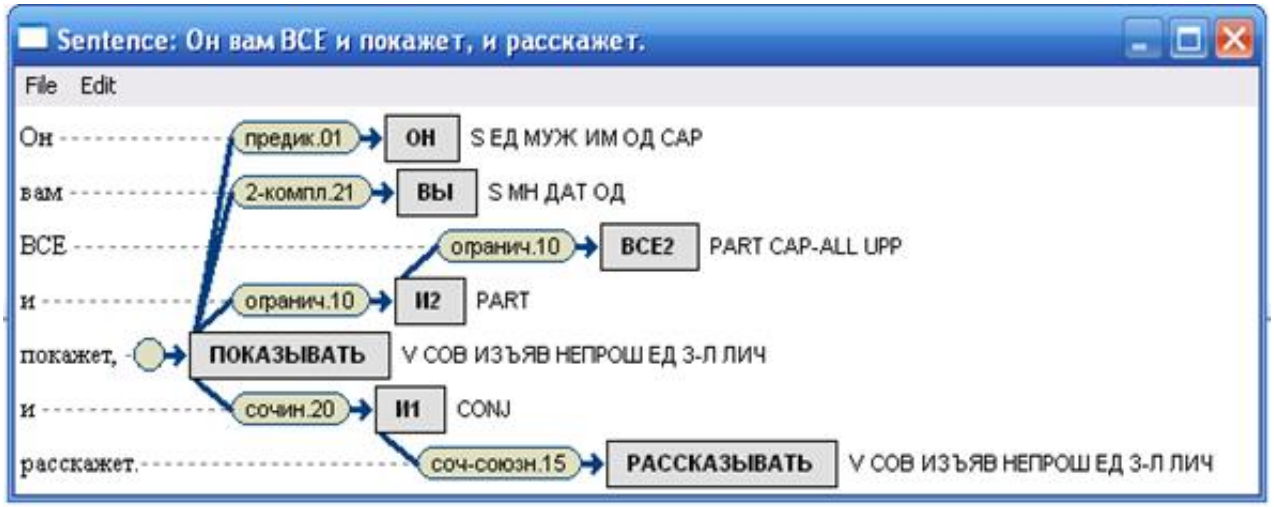

Рис. 8. Пример 3 правильного синтаксического разбора предложения со словом ***ВСЁ***

На рис. 9 – 10 приведены примеры правильного синтаксического разбора предложения со словом ***ВСЕ.*** При правильном разборе омограф ***ВСЕ*** маркируется всегда как местоимение-существительное (А) множественного числа.

Рис. 9. Пример 1 правильного синтаксического разбора предложения со словом ***ВСЕ***

Рис. 10. Пример 2 правильного синтаксического разбора предложения со словом ***ВСЕ***

На рис. 11 и 12 приведены примеры неправильного синтаксического разбора предложения со словом *ВСЁ*. В этих примерах слово ***ВСЁ*** ошибочно распознано как ***ВСЕ*** , т.е. как местоимение-прилагательное (рис.11), либо как местоимение-существительное (рис.12) множественного числа.

Рис. 11. Пример 1 неправильного синтаксического разбора предложения со словом ***ВСЁ***

Рис. 12. Пример 2 неправильного синтаксического разбора предложения со словом ***ВСЁ***

В заключение заметим, что при использовании системы ЭТАП-3 на всём протестированном тексте (роман Б. Акунина «Азазель»), в котором присутствовало 123 вхождения омографа ***ВСЕ***, обнаружено лишь 5 ошибочных отнесений слова ***ВСЁ*** к слову ***ВСЕ***, т.е. только 4% ошибочного распознавания!

Заключение

Однозначного ответа на вопрос, поставленный в качестве эпиграфа к этой статье, пока не существует. Однако, с уверенностью можно сказать, что полное алгоритмическое решение задачи расстановки недостающих точек над «Ё» наступит не ранее, чем в полной мере будут решены проблемы морфологического, синтаксического, семантического и прагматического анализа текстов. Например, как понять: *ВСЁ ДЕРЬМО,* или *ВСЕ ДЕРЬМО?* Система «ЭТАП» говорит, что *ВСЁ*.

В заключение хочу выразить искреннею благодарность ***Елене Ягуновой*** за предоставление словаря омографов [2] и за подсказку использовать в работе Интернет ресурс [3]. И, наконец, но не в последнюю очередь, ***Леониду Иомдину*** за предоставленную мне возможность использования синтаксического анализатора «ЭТАП-3» в ходе выполнения данной работы.

Список литературы

1. Д.Э. Розенталь, М.А. Теленкова. Словарь-справочник лингвистических терминов // Изд. «Просвещение», М. 1976, 543 с..
2. А.В. Венцов и др. Словарь омографов русского языка // Изд. СПбГУ, Санкт-Петербург, 2004, 160 с.
3. Национальный корпус русского языка “Поиск по акцентуированному корпусу” // Интернет ресурс: http://www.narusco.ru
4. И.М. Богуславский, Л.Л.Иомдин, Д.Р. Валеев, В.Г. Сизов. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов // Труды Международной конференции <Корпусная лингвистика -2008>. СПб.: Санкт-

Петербургский государственный университет, 2008. С. 56-74.