Ягунова Е.В. Ключевые слова в исследовании текстов Н.В. Гоголя // Проблемы социо- и психолингвистики. Пермь, 2011 (в печати)

***Е.В. Ягунова** (Санкт-Петербург)*

Ключевые слова в исследовании текстов Н.В. Гоголя

Введение

В этой статье приводятся результаты исследования процедур понимания текста, где понимание текста приравнивается к извлечению из него информационной структуры. Извлечение наиболее важной информации, передаваемой текстом, **может** быть смоделировано через процедуры выделения ключевых слов текста. В результате этих процедур исследователь получает наборы ключевых слов (см., напр., (Сахарный и др. 1984; Сахарный и др. 1988; Сиротко-Сибирский 2006 и др.)). Ключевые слова в наборе заведомо неравноправны не только по степени уверенности отнесения слова к ключевым, но и по определению специфической роли для каждого ключевого слова.

Важно, что информационная структура текста принципиально неоднозначна, как неоднозначно наше понимание любого текста. Особенно ярко неоднозначность информационной структуры проявляется в случае моделирования художественного текста. В предлагаемом вашему вниманию модельном исследовании процедур понимания текста большое внимание уделяется формальным признакам и сопоставительному анализу поведения, условно говоря, разных типов адресатов: носителя языка и автомата.

С другой стороны, в этой статье приводится краткий анализ особенностей произведений Н.В.Гоголя с помощью предложенной методики. Причина выбора материала лежит в наличии тематически более или менее однородных серий произведений, позволяющих формировать миниколлекции для вычислительного эксперимента, и «тематической уникальности» каждой из этой миниколлекций.

Цели. Задачи

Цель нашего проекта в целом: изучить в ходе единого исследования разнообразные процедуры извлечения ключевых слов из текста, объединив существующие подходы и методики – от эксперимента с информантами до вычислительного эксперимента (см. об этом Ягунова 2010а; Ягунова 2010б) . Один из важных аспектов этого проекта: анализ зависимости реализуемых процедур выделения ключевых слов:

- от функционального стиля текста (художественный, научный, новостной, официально-деловой),
- от темы, стиля, жанра и т.д.,
- от стиля конкретного писателя,
- от тематики произведения или цикла произведений рассматриваемого писателя.

В статье мы ставим следующие цели:

- исследование информационных структур, представленных распределением ключевых слов на фоне неключевых (всех прочих), для произведений Н.В. Гоголя,
- анализ формальных признаков, важных для формирования наборов ключевых слов,
- сопоставление информационных структур и/или наборов ключевых слов при восприятии текста информантом и автоматом.

Ограничимся в данном исследовании такими формальными признаками:

1) частота встречаемости слова (класса слов) в конкретном тексте,
2) распределение слова (класса слов) по тексту:
    - равномерность,
    - для неравномерных – тяготение к началу / концу текста.

Сопоставление частоты встречаемости в тексте с частотой встречаемости по корпусу можно считать третьим формальным признаком. Кроме того этот формальный признак, как правило, является главным для систем

автоматического понимания текста, т.е. для моделирования понимания таким адресатом, как автомат.

Для того чтобы были реализованы цели исследования, необходимо решить следующие **задачи**:

1) выделение ключевых слов в ходе вычислительного эксперимента с использованием коэффициента важности tf-idf,
2) выделение ключевых слов в ходе эксперимента с информантами,
3) сопоставительный анализ;
4) анализ распределения слов по тексту.

Материал и методика

Материал – 3 тематически наиболее однородные коллекции: 1) «Петербургский цикл», 2) «Мертвые души», 3) «украинская тематика»: «Миргород» и «Вечера на хуторе близ Диканьки»[1].

Основой для проведения вычислительного эксперимента служила **методика** с использованием меры TF-IDF; это традиционная статистическая мера, применяемая для оценки важности слова в контексте документа, являющегося частью коллекции документов[2]. Мера TF-IDF является произведением двух сомножителей: TF и IDF.

TF (*term frequency* — частота слова) оценивает важность слова t_i в пределах отдельного документа:

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k},$$

где n_i есть число вхождений слова в документ, а в знаменателе — общее число слов в данном документе.

[1] I. «Петербургские повести»: «Портрет», «Шинель», «Нос», «Невский проспект», «Коляска», «Записки сумасшедшего»; II. «Мертвые души»; III. Украинская тематика: «Вечера на хуторе близ Диканьки», и цикл «Миргород» («Вий», «Тарас Бульба», «Повесть о том, как поссорился Иван Иванович с Иваном Никифоровичем»).

[2] Программная реализация вычислительного эксперимента осуществлена Л.М.Пивоваровой. Пользуясь случаем, хочу выразить ей благодарность.

IDF (*inverse document frequency* — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт IDF уменьшает вес широкоупотребительных слов:

$$\mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|},$$

где |D| — количество документов в корпусе;

$|(d_i \supset t_i)|$ — количество документов, в которых встречается t_i (когда $n_i \neq 0$).

Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

На основании весов слов – значений меры – мы можем определить (потенциально) ключевые слова: слова, наиболее важные для содержания конкретного текста и/или подколлекции по отношению к заданному контексту (коллекции).

Правильность этого определения зависит, главным образом, от того, насколько правильно определен **контекст**, а именно коллекция, с которой сравниваются слова интересующего нас текста или подколлекции (цикла).

Анализируемые подколлекции (циклы) текстов Н.В. Гоголя сопоставлялись с оставшимися подколлекциями текстов Н.В. Гоголя и коллекциями текстов А.П.Чехова, сборниками «Человек в футляре», «Рассказы 1887 год», «Рассказы. Повести. 1888-1891», «Рассказы. Повести. 1892-1894», «Рассказы. Повести. 1894-1897»[3]. Наш выбор контекста определяется требованием максимальной однородности.

Для проведения эксперимента с информантами традиционную **методику** проведения эксперимента с информантами со стандартной инструкцией А.С. Штерн (Мурзин, Штерн 1991): Вспомните «Петербургские повести» Н.В.Гоголя. Подумайте над их содержанием. Выпишите 10-15 слов, наиболее важных для их содержания. И далее также – «Вспомните «Мертвые души»

[3] А.П. Чехов. Полное собрание сочинений и писем в 30-ти томах. Сочинения. Том 1. М., "Наука", 1983

Н.В.Гоголя. …» , «Вспомните украинский цикл Н.В.Гоголя («Миргород» и «Вечера на хуторе близ Диканьки»)...». Единственное отличие от традиционного варианта заключалось в том, что информантам предлагалось вспомнить тексты, т.е. оценивалось остаточное знание текста. В экспериментах по определению КС участвовало по 21 информанту для каждого из трех циклов.. В качестве информантов выступали профессиональные филологи (не студенты), хорошо знающие русскую классику. К участию в эксперименте не привлекались преподаватели русской литературы в школе или ВУЗе, чтобы образовательные методики, программы, стандарты не влияли на результат эксперимента.

Результаты

В таблице 1 приведены потенциально ключевые слова, выделенные с использованием коэффициента важности TF-iDF, слова упорядочены по убыванию значения этой меры. Пороговое значение определялось эмпирически.

Таблица 1 а-в. Ключевые слова, полученные в результате вычислительного эксперимента

а. Петербургские повести

Акакиевич	рука	**шинель**
Ковалев	лицо	ростовщик
Акакий	медж	асессор
Яковлевич	пуф	коллежский
маиор	**нос**	титулярный
Шиллер	квартальный	коломна
Чартков	бакенбарды	лорнет
Пискарев	время	прыщик
проспект	департамент	Рафаэль
Чертокуцкий	голова	Фидель
чорт	комната	Психея
портрет	**художник**	происшествие
человек	слово	**чиновник**
Невский	Испания	дама
глаза	штаб-офицерша	казаться

Гофман	беспрестанный	

б. Мертвые души

Чичиков	Копейкин	герой
Ноздрев	Мураз	**души**
Манилов	Антонович	дама
Селифан	Петрушка	голова
Собакевич	**бричка**	Леницын
Костанжогло	Платонов	поэма
человек	лицо	чубарый
Плюшкин	**купчая**	думать
Платон	Павел	Иванович
Хлобуев	город	жизнь
тентетник	сторона	Бог
слово	глаз	дом
рука	Кошкарев	барин
тентетников	место	полицеймейстер
время	ассигнация	председатель

в. Украинская тематика

козак	**Днепр**	человек
Никифорович	дьяк	лях
пан	**черевички**	**Вакула**
хата	рука	Миргород
запорожец	**Чуб**	**Солоха**
козацкий	**кузнец**	Хома
Андрий	свитка	есаул
Тарас	**Голова**	**панночка**
Остап	галушка	Григориевич
Данило	Левко	куренной
курень	Оксана	Прокофиевич

Катерина	Янкель	гетьман
Иван	хлопец	Дорош
Бульба	Петро	комиссар
парубок	сотник	**Иванович**
		шинок

В общем и целом, можно сказать, что определяемые таким образом слова представляют собой наименования действующих лиц, мест и событий. Полужирным шрифтом выделены слова, относящиеся к пересечению множеств ключевых слов, выделяемых в ходе вычислительного эксперимента (см. табл. 1) и в ходе эксперимента с информантами (табл. 2).

Для вычислительного эксперимента имеют существенное значение такие факторы, как частотность слова в тексте, наличие/отсутствие очевидной внутренней формы (напр., Коробочка) и даже «правильность» формоизменительной парадигмы.

В таблице 2 приведены результаты эксперимента с 21 информантом по выделению ключевых слов, количественные данные приведены в абсолютных числах (указывается число информантов, записавших в анкете данное слово).

Таблица 2 (а, б, в). *Ключевые слова, полученные в результате эксперимента с информантами*

а. Петербургские повести,

п.п	ключевые слова	i1	i2
1	**шинель**	15	15
2	**нос**	9	12
3	**художник**	11	11
4	**чиновник**	11	11
5	**Невский**	9	9
6	**Акакий**	9	9
7	**проспект**	8	8
8	**портрет**	6	8
9	сумасшествие	7	7
10	Петербург	7	7
11	мечта	4	5
12	**майор**	5	5
13	страх	4	4
14	холод	4	4
15	**Акакиевич**	3	3
16	обман	2	3

17	Пирогов	3	3

18	**Пискарев**	3	3

б. Мертвые души

п.п	ключевые слова	i1	i2	i3
1	помещик	5	10	
2	**бричка**	8	8	
3	тройка	8	8	
4	**Чичиков**	8	8	
5	дорога	7	7	
6	Коробочка	7	7	
7	**Плюшкин**	7	7	
8	**купчая**	6	6	7

9	**Манилов**	6	6	
10	**Собакевич**	6	6	
11	**души**	3	6	
12	мертвые	6	6	
13	губернатор	2	5	
14	**Ноздрев**	5	5	
15	крепостные	3	3	4
16	Россия	3	3	6

в. Украинская тематика

п.п.	ключевые слова	i1	i2	i3
1	черт	9	11	
2	ночь	9	9	12
3	**панночка**	7	7	
4	**кузнец**	7	7	
5	**черевички**	6	6	
6	Рождество	6	6	
7	любовь	5	5	
8	гусак	5	5	
9	**Иван Иванович**	5	5	
10	ярмарка	5	5	
11	ведьма	3	4	
12	**Голова**	4	4	
13	**Иван Никифорович**	4	4	

15	праздник	4	4	
16	**Солоха**	4	4	
17	**Чуб**	4	4	
18	**казак**	2	4	
19	парубок	2	4	
20	Украина	3	4	
21	Ивана Купала	2	3	
22	**Вакула**	3	3	
23	вий	3	3	
24	**Днепр**	3	3	
25	еда	3	3	
26	звезды	3	3	
27	нечисть	3	3	
28	русалка	2	3	

29	смех	3	3		30	хутор	3	3	

Условные обозначения: «i1» -- число информантов, записавших слово в данной форме, «i2» -- число информантов, записавших данную лексему, «i3» -- число информантов, записавший слово с точностью до «класса эквивалентности» (напр., *губернатор* и *губернаторский*; *Россия* и *Русь*).

По предварительным результатам информационная структура «Петербургских повестей» отличается максимальной компактностью и прозрачностью; для этого цикла наблюдается достаточно большое количество ключевых слов, выделяемых на основании и эксперимента с информантами, и вычислительного эксперимента (см. табл. 1а и 1б). Списки потенциально ключевых слов, выделяемых на основании вычислительного эксперимента и эксперимента с информантами (см. табл.1а и табл.2а), хорошо демонстрируют различия между двумя типами информационных структур: извлекаемой человеком в процессе понимании текстов vs. автоматом при реализации процедур информационного поиска.

Информационная структура подколлекции «украинская тематика» характеризуется максимальной неоднородностью. Обращает внимание то, что списки ключевых слов, выделяемые для этой подколлекции в ходе вычислительного эксперимента, интуитивно кажутся адекватными для понимания текстов носителем языка (представления информационной структуры носителями языка).

Данные, полученные на материале поэмы «Мертвые души», оказываются промежуточными (между подколлекциями «Петербургские повести» и «украинская тематика»).

Обсуждение результатов

Слова, являющиеся «символами текста», далеко не всегда могут определяться в ходе вычислительного эксперимента. Например, лексема «тройка» (в частности, «*Эх, тройка! птица тройка, кто тебя выдумал? знать, у бойкого народа ты могла только родиться…*») встречается 13 раз в тексте;

однако, вряд ли кто-нибудь усомнится в значимости этого ключевого слова для нашего представления о тексте «Мертвые души» (38% информантов записало это слово в своей анкете). Слово «дорога» оказывается ключевым по мнению информантов (опять же 38% информантов его записало в анкетах). Однако это слово является довольно частотным в русском языке (и, в частности, в текстах Н.В. Гоголя и А.П. Чехова, составляющих контрастивную коллекцию). В «Мертвых душах» эта лексема встречается 119 раз, но в вычислительном эксперименте это слово не было выделено в качестве ключевого, ведь оно частотно не только для этого произведения.

Аналогичными примерами слов, являющимися ключевыми только на основании эксперимента с информантами, являются следующие:

- «Петербургские повести» – *мечта* (10), *обман* (2), *страх* (20), *холод* (2);
- «Украинская тематика» – *ночь* (125), *любовь* (10) [4].

Проиллюстрируем анализ роли таких формальных признаков как частота встречаемости в конкретном тексте и распределение по тексту.

Для того чтобы подобный анализ был наиболее наглядным, Д. Ландэ были «реализованы инструментальные средства, позволяющие визуализировать плотность встречаемости слова в тексте в зависимости от ширины окна наблюдения. В … спектрограмме по горизонтали откладываются номера вхождения слова в тексте, а по вертикали – ширина окон наблюдения (начиная со значения 1 в самом низу, вхождения слова в данном случае выделяется светло-серым цветом). Если в соответствующее окно наблюдения попадает несколько целевых слов, то оно закрашивается более интенсивным оттенком темного» (Ландэ 2009)[5].

Возьмем для примера некоторые ключевые слова и их формальные описания на материале первого тома «Мертвых душ». Наиболее

---

[4] В скобках приведена частотность лексемы в рассматриваемом тексте и/или цикле (коллекции)

[5] Сервис Д.В. Ландэ доступен по адресу http://ling.infostream.ua/jag/

иллюстративным оказывается описание разных действующих лиц (ср. Ягунова 2010б).

На рис. 1 представлена спектрограмма, отражающая распределение наименования главного действующего лица: лексема «Чичиков» в тексте встречается 467 раз.

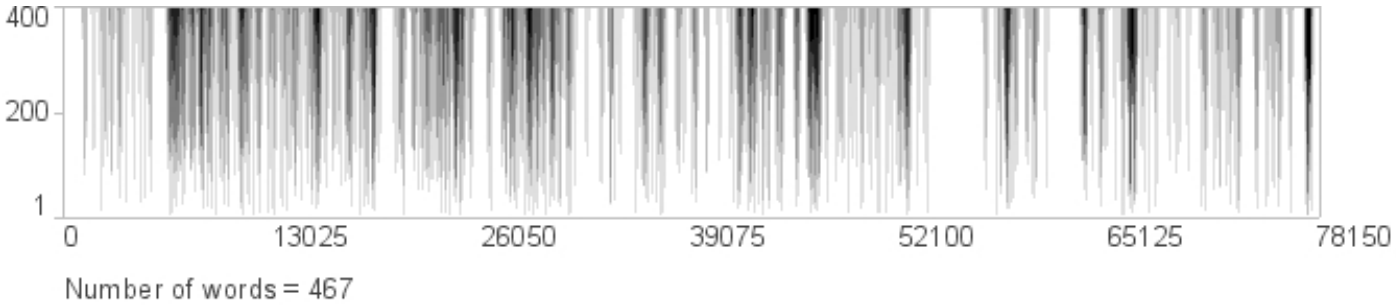

Рис. 1. Чичиков

Интерпретация очень высокой частоты встречаемости и почти равномерного распределения по тексту с большой вероятностью – уже на основании анализа формальных признаков – приводит к решению о том, что «Чичиков» является главным действующим лицом (ср. табл. 1б и 2б).

Покажем возможность классификации других действующих лиц. На рис. 2 представлена спектрограмма для лексемы «Манилов» (105 словоупотреблений в тексте (с/у)), на рис. 3 – для лексемы «Ноздрев» (143 с/у), на рис. 4 – для лексемы «Собакевич» (106 с/у), на рис. 5 – для лексемы «Плюшкин» (46 с/у), на рис. 6 – для лексемы «Копейкин» (32 с/у), на рис. 7 – для лексемы «помещик» (44 с/у).

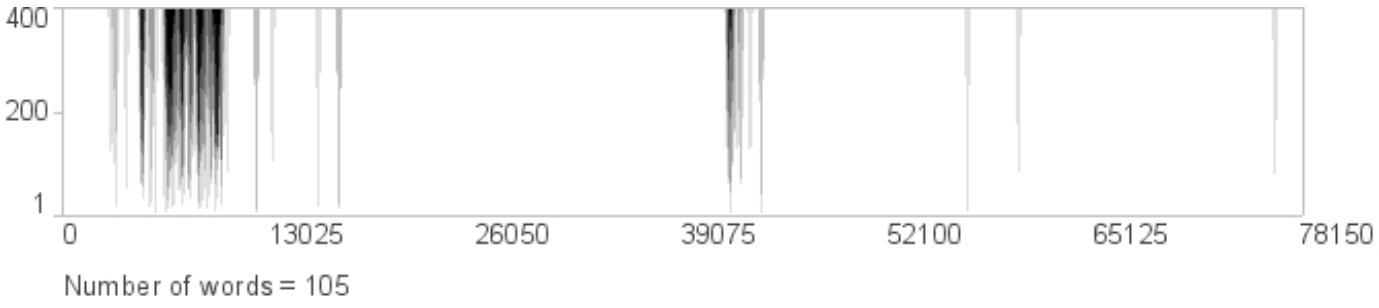

Рис. 2. Манилов

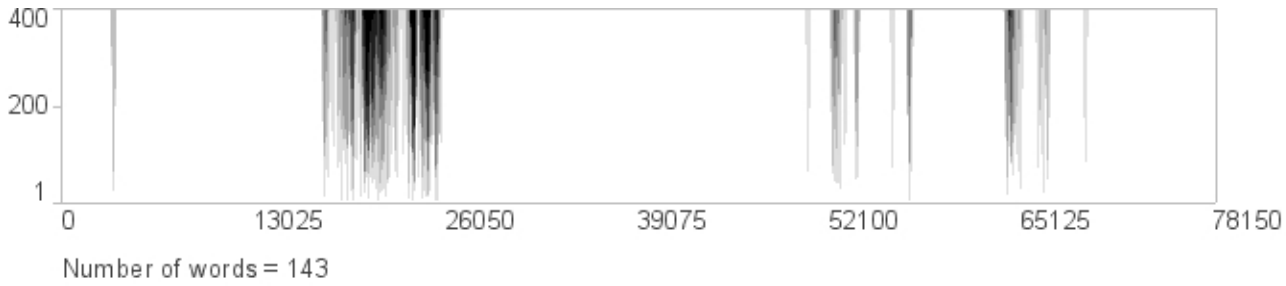

Рис. 3. Ноздрев

Лексемы «Манилов» и «Ноздрев» характеризуют следующие признаки:

- высокая частота встречаемости;
- неравномерность распределения, т.е. они сосредоточены главным образом в «своем» фрагменте текста, для этих двух лексем наблюдается почти полное отсутствие пересечения.

Лексема «Манилов» тяготеет к началу текста, что, вероятно, сообщает ей несколько более высокую степень «важности», которая может компенсировать чуть меньшую частоту встречаемости (по сравнению с лексемой «Ноздрев»).

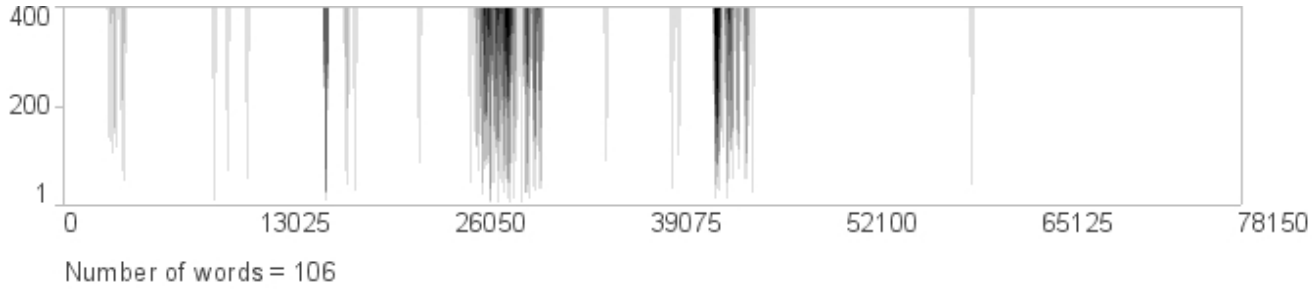

Рис. 4. Собакевич

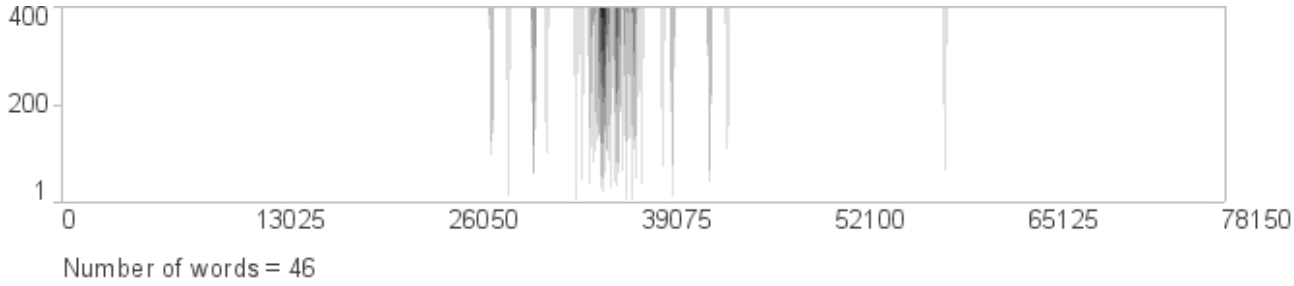

Рис. 5. Плюшкин

Лексему «Собакевич» (как и «Манилов», «Ноздрев») характеризует высокая частота встречаемости, но ее отличает рассредоточение по двум основным фрагментам. Лексему «Плюшкин» отличает средняя частота встречаемости и сосредоточение, главным образом, на одном фрагменте.

Между распределениями лексем «Собакевич» и «Плюшкин» наблюдается пересечение.

Все рассмотренные выше ключевые слова выделяются как в вычислительном эксперименте, так и в эксперименте с информантами.

Дополнить классификацию действующих лиц нам поможет рассмотрение таких специфичных потенциально ключевых слов, как лексемы «Копейкин» и «помещик». Потенциально ключевое слово «Копейкин» выделяется только в вычислительном эксперименте (см. табл.1б), а «помещик» определяется только в эксперименте с информантами (причем именно «помещик» записывает в анкетах большинство информантов, см. табл. 2б).

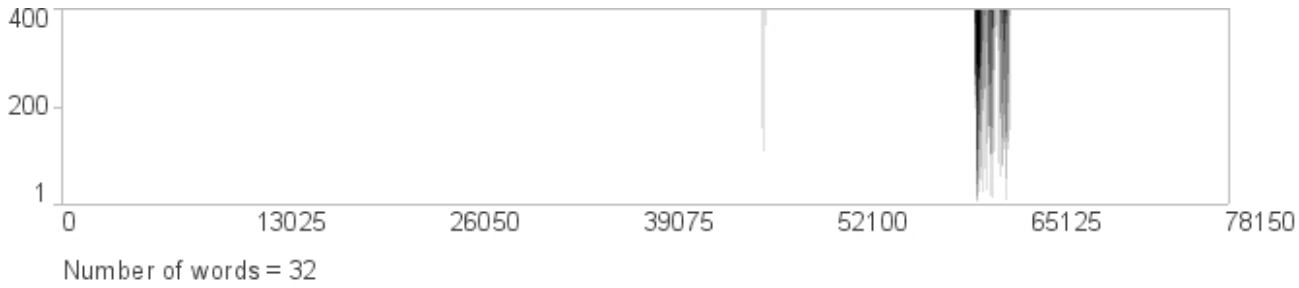

Рис. 6. Копейкин

Для лексемы «Копейкин» характерна средняя частота встречаемости, она сосредоточена только на одном фрагменте («Повесть о капитане Копейкине»), на основании формальных критериев этому слову можно было бы приписать сравнительно высокую степень «важности».

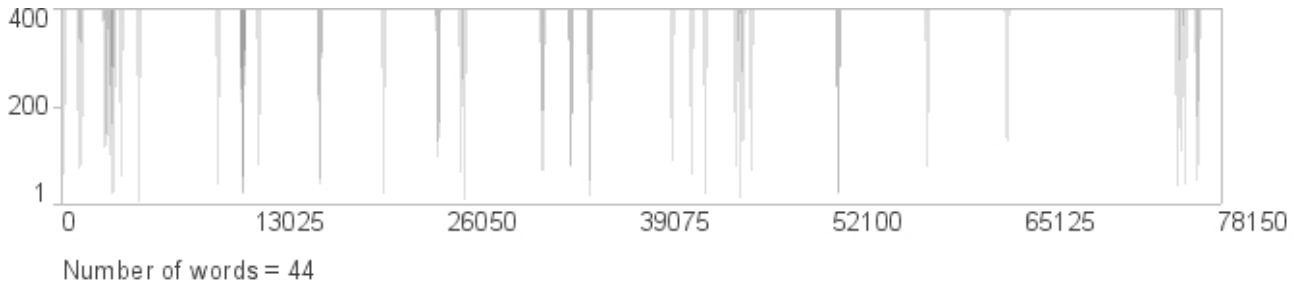

Рис. 7. Помещик

Лексеме «помещик» свойственна средняя частота встречаемости и почти равномерное распределение; на основании формальных критериев вероятна низкая степень «важности». Обобщенность этой номинации делает ее как бы малоинформативной как с точки зрения рассматриваемых формальных

признаков, так и с точки зрения возможностей вычислительного эксперимента.

Второй тип рассматриваемых нами ключевых слов, можно назвать «ключевые слова, аккумулирующие содержательные вехи описания и/или рассуждения» (ср. Ягунова 2010б).

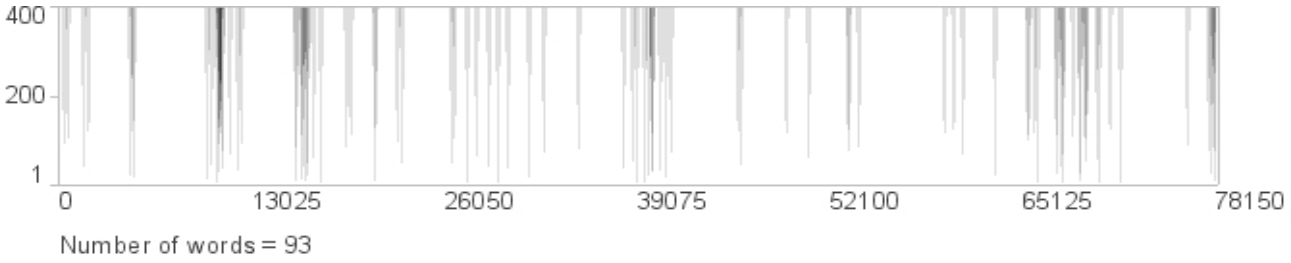

Рис. 8. Дорога

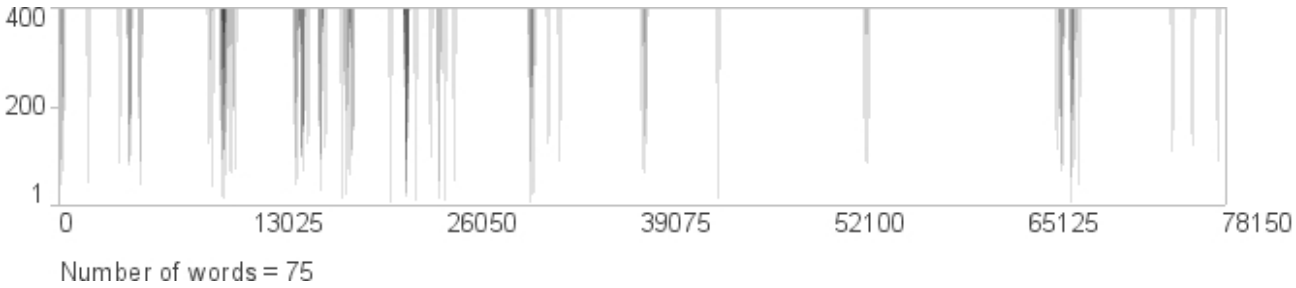

Рис. 9. Бричка

Лексемы «дорога» и «бричка» характеризуют средневысокая частота встречаемости, сравнительно равномерное распределение по тексту (включая конец или начало текста). Эти признаки дают возможность, рассматривать лексемы «дорога» и «бричка» как потенциально ключевые. Лексема «бричка» оказывается ключевой на основании всех данных (ср. табл. 1б и 2б). Лексема «дорога» не выявляется лишь на основании вычислительного эксперимента в силу высокой общеязыковой частоты встречаемости. С некоторой долей условности можно сказать, что эти два ключевых слова ведут себя как аналоги наименований действующих лиц.

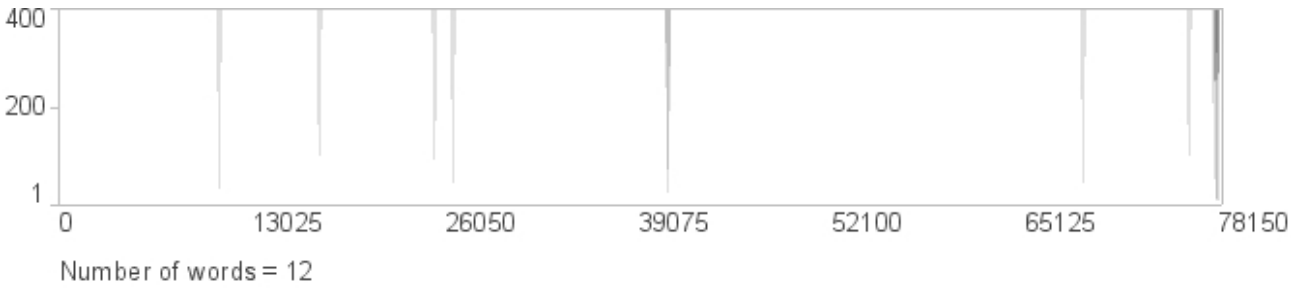

Рис. 10. Тройка

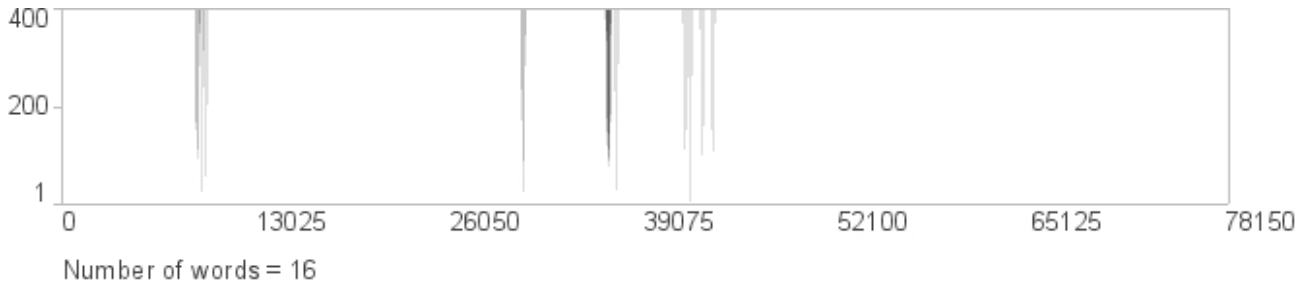

Рис. 11. Купчая

Лексема «тройка» выделялась только в ходе эксперимента с информантами; и лексема «тройка», и лексема «купчая» характеризуются низкой частотой встречаемости.

В распределении потенциального ключевого слова «тройка» есть сосредоточение в конце текста, поэтому все же есть возможность формально отнести его к разряду важных.

В распределении потенциального ключевого слова «купчая» нет тяготения к началу или концу текста, невозможно на основании данных о распределении отнести ее к «важным». Эта лексема определяется как ключевая в ходе обоих экспериментов, статистическая мера TF-iDF оказывается эффективной благодаря низкой частоте встречаемости (даже для анализируемого контекста).

Вместо заключения

Как уже говорилось во введении, информационная структура текста – особенно художественного текста – принципиально неоднородна и неоднозначна как неоднозначно наше понимание любого текста. Обсуждаемые в статье наборы ключевых слов, полученных в ходе вычислительного эксперимента и эксперимента с информантами, дают возможность сопоставительного анализа разных видов понимания. В качестве результатов понимания мы рассмотрели информационные структуры, извлекаемые принципиально разными типами адресатов: носителем языка и автомата.

С другой стороны в этой статье были кратко показаны основные различия информационных структур в зависимости от выбора одной из трех коллекций («Петербургские повести», «Мертвые души» и «украинская тематика»). Они

различаются как степенью однородности, так и преобладающими типами ключевых слов.

На материале «Мертвых душ» представлена иллюстрация возможностей использования формальных признаков (видов распределения в тексте) и варианта классификации типов ключевых слов (прежде всего, действующих лиц).

Для полноты картины сопоставим полученные результаты с теми данными, что были получены нами на материале научных текстов (предметной области "корпусная лингвистика"). Для научных текстов предлагаемая методика дает еще более четкие результаты выделения и классификации ключевых слов (Ягунова 2010). Различие между художественными и научными текстами состоит, прежде всего, в весах этих признаков. В частности, различительная сила слова, оцениваемая с использованием третьего формального признака (Tf-iDF), гораздо выше для научного текста, чем для художественного.

В результате предложенного подхода можно получить ответы на многие вопросы, рассматривая их как следствие различий информационных структур текстов (коллекций):

- о различии между текстами внутри одного функционального стиля,
- о различии между текстами (информационными структурами) в зависимости от функционального стиля текста,
- о ядерном или периферийном положении текста в рассматриваемом корпусе или корпусах и т.д.

Литература

1. Ландэ Д.В. Визуализация статистики вхождения слов // MegaLing'2009. Горизонты прикладной лингвистики и лингвистических технологий. Материалы международной конференции 21-26 сентября 2009 г., Украина, Киев 2009. – С. 63-64

2. Мурзин Л. Н., Штерн А. С. Текст и его восприятие.– Сверд-ловск : Изд-во Урал. ун-та, 1991. – 172 с.

3. Сахарный Л. В., Сибирский С. А., Штерн А. С. Набор ключевых слов как текст // Психолого-педагогические и лингвистические проблемы исследования текста. – Пермь, 1984. – С. 81-83.

4. Сахарный Л. В., Штерн А. С. Набор ключевых слов как тип текста // Лексические аспекты в системе профессионально-ориентированного обучения иноязычной речевой деятельности. – Пермь, 1988. – С. 34—51.

5. Сиротко-Сибирский С. А. О проблеме понимания текста в лингвистике и психолингвистике // … СЛОВО ОТЗОВЕТСЯ : памяти Аллы Соломоновны Штерн и Леонида Вольковича Сахарного / Перм. ун-т. – Пермь, 2006. – С. 63-68.

6. Ягунова Е.В. Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь, 2008

7. Ягунова Е.В. Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов // Материалы IV Международного конгресса исследователей русского языка «Русский язык: исторические судьбы и современность». М., 2010а. – С. 533-534

8. Ягунова Е.В. Эксперимент и вычисления в анализе ключевых слов художественного текста // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика / Отв. ред. В.Т. Юнгблюд. Вып. 1. – Пермь, 2010б. С. 85-91.