Е.В. Ягунова Эксперимент и вычисления в анализе ключевых слов художественного текста // Философия языка. Лингвистика. Лингводидактика №1 Пермь 2010. с.83-89

***Е.В. Ягунова** (Санкт-Петербург)*

Эксперимент и вычисления в анализе ключевых слов художественного текста

В статье излагаются пути выделения и классификации ключевых слов художественного текста. Методика предполагает объединение традиционного эксперимента с информантами и анализа формальных признаков, значения которых выявляются в ходе вычислительного эксперимента.

Введение

Моделирование восприятия и понимания текста человеком, определение того, на основании каких критериев он принимает решение о том, что является смысловыми вехами текста, является актуальным и значимым для теории восприятия и понимания текста. Несмотря на существование многих моделей, процедуры извлечения информации из текста, моделируемые в нашем проекте через процедуры выделения ключевых слов, остались практически неизученными.

В современном информационном обществе большое внимание уделяется созданию автоматических систем понимания текста, извлечению информации как из отдельно взятого текста, так и из информационных потоков (множества однотипных текстов). За право на «свое понимание» понятия ***ключевое слово*** сражаются представители разных областей знаний. Пожалуй, наиболее ярко это противостояние можно проследить, сопоставляя работы о ключевых словах психолингвистической парадигмы и интенсивно развивающейся области информационного поиска (или – хотя бы – заглянув в википедию).

Согласимся, что извлечение наиболее важной информации, передаваемой текстом, ***может*** быть смоделировано через процедуры выделения ключевых слов (КС) текста. В результате этих процедур исследователь получает наборы КС (см., напр., (Сахарный и др. 1984; Сахарный и др. 1988; Сиротко-Сибирский 2006 и др.)). КС в наборе заведомо неравноправны не только по степени

уверенности отнесения слова к ключевым, но и по определению специфической роли для каждого КС.

Цели. Задачи

В рамках исследования проблемы ключевых слов предлагаются подходы для решения вопроса о том, ***как*** происходит определение КС носителем языка:

а) на основании каких признаков происходит определение КС;

б) на каком фрагменте текста происходит (может происходить) это определение, как оно соотносится с необходимостью подстройки адресата под смысловые особенности текста (см. Ягунова 2008);

в) как могут быть описаны типы КС.

Решение этого вопроса может заключаться в сопоставлении результатов выделения КС на основании эксперимента с носителями языка и вычислительного эксперимента.

В качестве исследуемых формальных признаков на данном этапе рассматриваются следующие:

1. частота встречаемости слова (класса слов) в конкретном тексте,
2. распределение слова (класса слов) по тексту:
    - равномерность,
    - для неравномерных – тяготение к началу / концу текста,
3. сопоставление частоты встречаемости в тексте с частотой встречаемости по корпусу (ср. коэффициенты уникальности слова, напр., TF-iDF[1], традиционно используемый в информационном поиске для оценки различительной силы слова текста).

Ранее были поставлены многие из сформулированных положений (см., напр., Ягунова 2008), однако для «доказательной базы» исследования необходимо сформировать (1) базовые выборки текстов для каждого из функциональных стилей и (2) однородные корпуса для уже сформированных

[1] Название статистической меры TF-iDF происходит от англ. TF — term frequency, IDF — inverse document frequency. Эта мера является произведением двух сомножителей: TF и IDF. Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

выборок. Для проекта в целом основным исходным материалом являются художественные (прозаические, сбалансированные по времени создания), научные (предметной области "корпусная лингвистика") и новостные тексты. Базовая выборка содержит по 10 текстов для каждого функционального стиля (типа). По результатам анализа базовая выборка может расширяться. Требование однородности корпуса текстов и сбалансированности по отношению к базовой выборке текстов является важным условием для адекватного учета третьего из формальных признаков. Различительная сила оценивается относительно корпуса текстов того же типа, а не относительно контрастивного корпуса.

Материал и методика

В данной статье рассматривается определение КС лишь на художественных текстах (нарративах). Это накладывает свои ограничения на возможности использования формальных признаков. Нами тщательно анализировались 10 художественных текстов. Небольшой объем статьи позволяет лишь кратко проиллюстрировать возможности применения нашего подхода на примере двух текстов: В. Астафьев «Ловля пескарей» (объемом ок. 10 тыс. словоупотреблений (с/у)) и В. Пелевин «Проблема верволка» (объемом ок. 11 тыс. с/у).

В экспериментах по определению КС участвовало по 21 информанту для каждого текста. Эксперименты проводились со стандартной инструкцией А.С. Штерн: «Прочитайте текст. Подумайте над его содержанием. Выпишите 10-15 слов, наиболее важных с точки зрения его смысла» (Мурзин, Штерн 1991).

Результаты

В таблицах 1а и 1б представлена общая информация о КС (в абсолютных числах). Число информантов, выделивших слово как ключевое, приводится в графе «лексема» части «КС»; данные о числе с/у интересующих нас слов – в

части «число с/у по тексту». Каждая часть упорядочена по убыванию этих значений.

В графе «тематич. класс» – присутствующей в обеих частях – приводятся данные для слов, объединенных в один класс единой или близкой тематикой (смыслом) применительно к конкретному тексту; чаще всего это разные номинации, относящиеся к одному объекту или классу объектов, напр., объединяются «*машина* + *ЗИЛ*», «*волк* + *верволк* + *вервольф*» или «*Грузия* + *грузины* + *грузинский*» и т.д.

В таблицах 1а и 1б полужирным шрифтом выделены слова, для которых ниже иллюстрируются типы распределения по тексту.

Таблица 1а. КС и число с/у по тексту: В. Астафьев «Ловля пескарей»

КС			число с/у по тексту		
	лек-сема	тематич. класс		лексема	тематич. класс
Грузия	15	20	**Отар**	54	54
Витязь	12	16	дом	31	34
гость(и)	9	14	земля	27	33
собор	7	7	брат	26	27
Гелати	7	10	гость(-и)	18	21
храм	7	7	русские	18	23
рыбалка	6	14	рыба	18	36
брат (-тья)	6	6	стол	18	22
земля	6	6	сердце	17	19
русский	6	6	человек	17	25
Дом	6	6	друг	16	23
гостеприимство	5		**Грузия**	15	34
грузин (-ы)	5		пескарей	15	
подарочек	5	5	Храм	15	15
праздник	5	9	собор	13	13
рыба	5		Гелати	12	13
философия	4	5	рак	11	11
человек	4	4	творчества	9	17
писатель	4	4	родник	8	11
Отар	4	4	**Витязь**	7	7
река	4	7	небо	6	9
небо		5	писатель	5	8
добросердечные		4	товарищ	5	
			реках	4	16

Таблица 1б. КС и число с/у по тексту: В. Пелевин «Проблема верволка»

КС			число с/у по тексту		
	лексема	тематич. класс		лексемы	тематич. класс
Саша	14	14	**волк**	35	38
дорога	10	10	эликсир	7	7
драка	9	11	страх	4	7
стая	9	9	стая	21	21
зов	9	9	совы	4	4
Коньково	9	9	**Саша**	192	205
машина	8	13	поляна	23	23
костер	8	8	ночь	8	26
деревня	8	8	**машина**	26	29
Вервольф	8		луна	11	12
волк	7	22	лес	26	28
поездка	7	7	Лена	27	27
лес	7	7	костер	19	19
ночь	7	7	**Коньково**	20	29
девочка	7	7	зов	8	8
поляна	7	7	дощатых	5	7
Лена	7	7	**дорога**	41	41
луна	7	7	деревня	9	
страх	6	6	девочка	11	15
эликсир	6	6			
совы	5	5			
оборотни	5	5			
ЗИЛ	5				

Д. Ландэ «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)[2]. На рис. 1-3 представлены спектрограммы для КС из В. Астафьева «Ловля пескарей», на рис. 4-6 – для КС из В. Пелевина «Проблема верволка».

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

Обсуждение результатов

Какие типы КС можно описать через набор этих признаков? Для всех рассматриваемых художественных текстов выделяется 4 основных типа КС:

1. Наименования главных действующих лиц (обычно самая высокая частота встречаемости, сравнительно равномерное распределение по тексту) – ср. рис. 1 и 4. Степень выраженности этих формальных признаков, как правило, соотносима со значимостью действующего лица и и/или степенью динамичности развития сюжета (ср. рис.1 vs. рис. 4). На нашем примере это одно из свидетельств большей значимости действующего лица («Саша») и большей динамичности сюжета в «Проблеме верволка» В. Пелевина по сравнению с «Отаром» в «Ловле пескарей» В. Астафьева.

2. КС, аккумулирующие содержательные вехи описания и/или рассуждения; эти КС могут вести себя аналогично наименованию действующих лиц – ср. рис. 2. «Грузия» в «Ловле пескарей» В. Астафьева по праву может быть названа одним из действующих лиц произведения; она была выделена большинством испытуемых как ключевое слово, характеризуется равномерным распределением по тексту и сравнительно частотно в тексте.

3. КС, отражающие основные идеи нарративного текста, соотносимые с начальным (преамбула и завязывание сюжета) или конечным (развязка) фрагментами – ср. рис. 5 и 6, соответственно. Наиболее ярко этот тип КС проявляется в нарративах с четкой структурой и динамичным развитием сюжета.

4. КС, имеющие сравнительно низкую частоту встречаемости по тексту, неравномерное распределение (напр., сосредоточены на одном композиционном фрагменте) – при крайне низкой общеязыковой частоте встречаемости (редкое слово) – ср. рис.3. Этот тип КС характерен скорее для научных или новостных текстов. Если четвертый тип КС встречается в рамках художественного текста, он позволяет выделить фрагмент, несущий особую роль в смысловой структуре текста (условно говоря, дополнительный

«центр тяжести»). «Витязь» в «Ловле пескарей» В. Астафьева – это, конечно, «Витязь в тигровой шкуре», олицетворение духа Грузии и знаковый композиционный фрагмент нарратива.

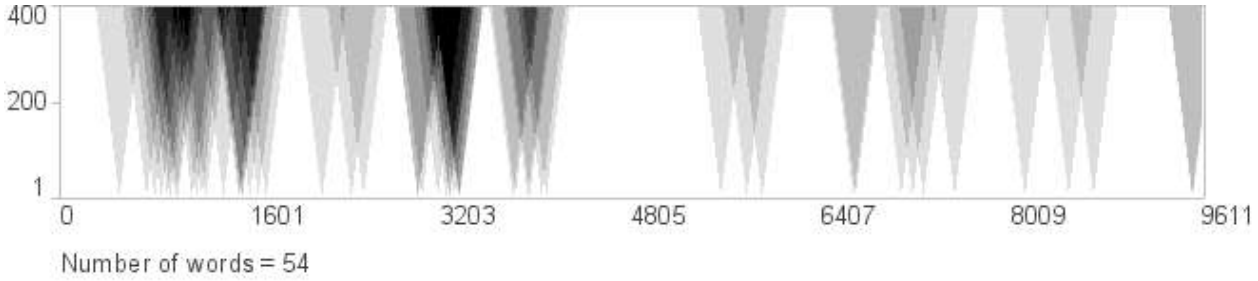

Рис. 1. *Спектрограмма вхождения слова «*Отар»

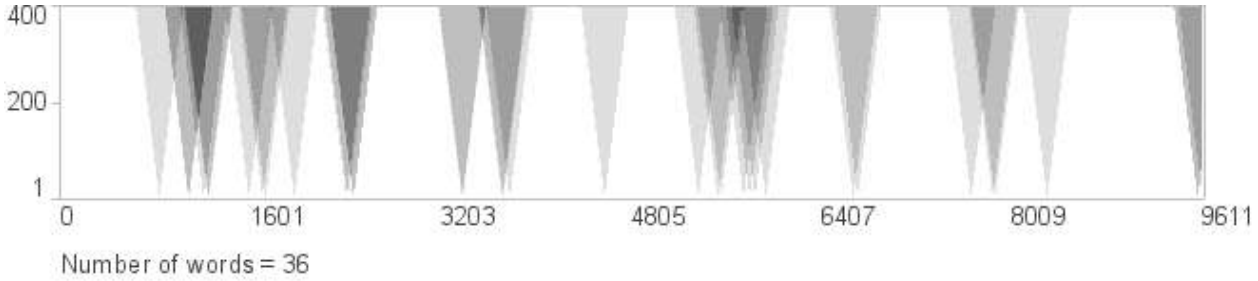

Рис. 2. *Спектрограмма вхождения слова «*Грузия»

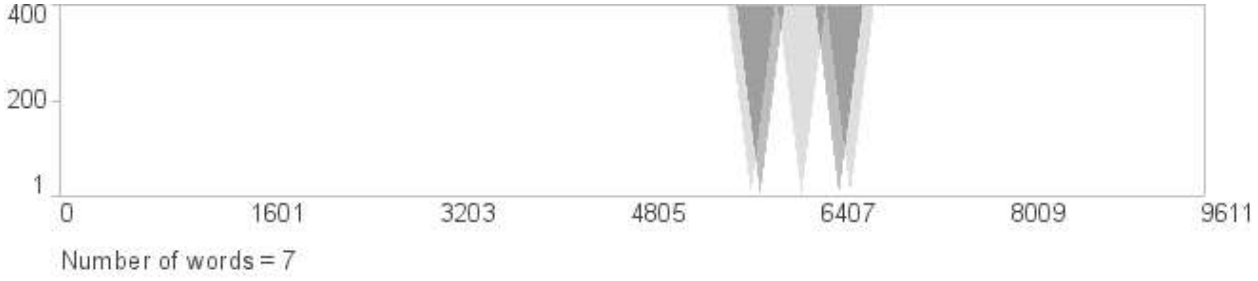

Рис. 3. *Спектрограмма вхождения слова «*Витязь»

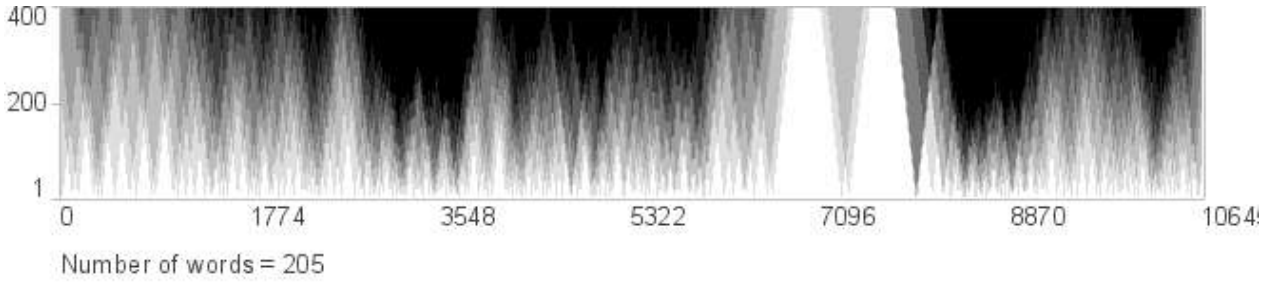

Рис. 4. *Спектрограмма вхождения слова «*Саша»

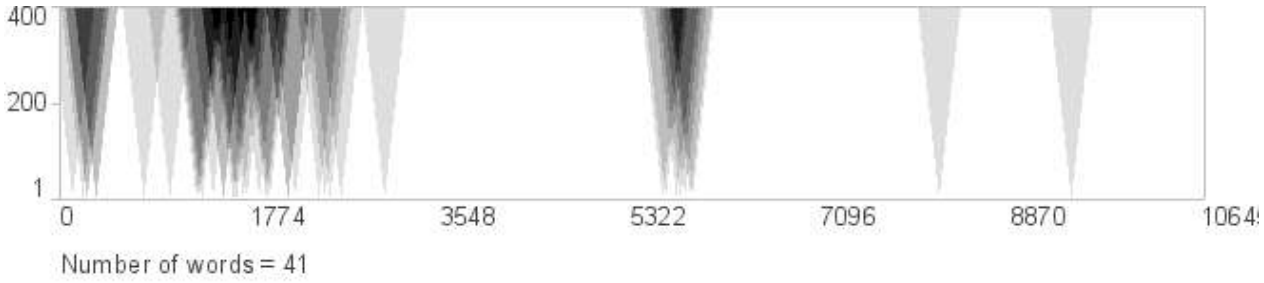

Рис. 5. *Спектрограмма вхождения слова «*дорога»

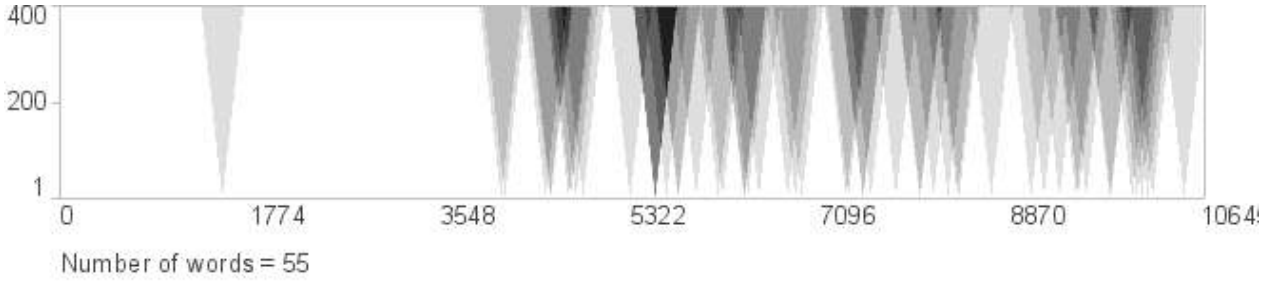

Рис. 6. *Спектрограмма вхождения слов «*волк, Верволк»

Тексты-примеры различаются степенью динамичности, что также прослеживается на основании формальных признаков (см. 1 и 3 типы КС).

Все приведенные типы КС далее на основании комбинации этих же формальных признаков распадаются на более дробные подтипы. Тип 1 – для текстов с несколькими ключевыми действующими лицами – может разбиваться на подтипы в соответствии со статусом этих действующих лиц и распределением по фрагментам текста. Само определение типа 3 дает возможность отнесения КС к подтипам этого класса в зависимости от тяготения к началу или концу текста. Дробность классификации можно продолжить.

Вместо заключения

Какие типы КС можно описать через набор этих признаков? Является ли предлагаемая классификация содержательной? Полагаем, что содержательность такой классификации очевидна. В результате выделяются типы и подтипы КС, оказывающиеся ведущими для структурной схемы нарратива.

Какую часть набора КС, выделенных информантами, можно описать через предлагаемый набор формальных признаков? Ответ на этот вопрос в существенной степени зависит от структуры текста. Для всех анализируемых художественных текстов (и, по-видимому, для любых художественных текстов

нарративной структуры) бóльшая часть КС может быть описана через описанные формальные признаки.

Как уже говорилось в начале статьи, в нашем исследовании рассматриваются тексты разных функциональных стилей. В завершение предложим некоторое обобщение.

Результаты, полученные на материале научных текстов (предметной области "корпусная лингвистика"), демонстрируют еще более четкие результаты классификации КС на основании тех же формальных признаков (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 («частотность слова в тексте / частотность слова в корпусе»), гораздо выше для научного текста, чем для художественного. С помощью предлагаемого подхода моделируется извлечение информационной (смысловой) структуры текста и описание этой структуры. В результате можно получить ответы на многие вопросы, рассматривая их как следствие различий информационных (смысловых) структур:

- 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 (с точки зрения структуры, предметной области и т.п.) 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., 2010. – С. 533-534