


Динамические частотные характеристики как основа для структурного описания разнородных лингвистических объектов

Ландэ Дмитрий Владимирович,
доктор технических наук, ИПРИ НАН Украины

Ягунова Елена Викторовна,
доктор филологических наук, С.-Петербургский гос. Университет

Переславль-Залесский,
15-18 октября 2012 года

Два определения для метода, базирующегося на сопоставлении частотных характеристик:

1. Глобальная частота встречаемости – абсолютная частота встречаемости слова в анализируемом объекте (от коллекции до текста).

2. Локальная частота встречаемости – абсолютная частота встречаемости слова в окне наблюдения из *K* слов.

В теории информационного поиска признано ранжирование слов по классическому критерию Солтона *TF IDF* [1], где *TF* (*Term Frequency*) – это частота встречаемости слова в пределах выбранного документа, а *IDF* (*Inverse Document Frequency*) – величина, обратная количеству документов, в которых встретилось данное слово.

Наш подход близок к *TF*, можно считать, что локальная частота – это аналог *TF* (в этом случае окно наблюдения – аналог документа), а глобальная частота встречаемости соответствует обратной *IDF*. При этом появляется возможность анализировать не только массивы документов, как это реализовано с помощью *TF IDF*, но и цельные тексты больших объемов (ср. [2]).

1. *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval. Information Processing and Management, 1988. – № 24(5). – P. 513-523.
2. *Ягунова Е.В.* Ключевые слова в исследовании текстов Н.В. Гоголя // Проблемы социо- и психолингвистики. Вып. 15. Пермь, 2011

В [3] исследовалась зависимость особенности соотношения локальной и глобальной популярности сообщений электронных СМИ. При этом было выявлено некоторое количество сообщений, характеризующихся большим соотношением локальной популярности к глобальной. Этот факт позволяет судить о событиях, описываемых в данных сообщениях, как о новых. Таким образом был обоснован алгоритм выявления документов, получивших большую популярность только в последнее время (*New Event Detection*) [4].

3. Ландэ Д.В., Григорьев А.Н., Брайчевский С.М., Дармохвал А.Т., Снарский А.А. Особенности соотношения локальной и глобальной популярности сообщений электронных СМИ // MegaLing'2007. – Симферополь, Изд-во: "ДиАйПи", 2007. - С. 223-224.

4. Ландэ Д.В., Фурашев В.Н. Выявление новых событий в рамках системы контент-мониторинга // Научно-техническая информация. – Сер. 2. Информационные процессы и системы. №12 – 2006. – С. 17-20.

Предлагаемый подход позволяет анализировать структуры самых разных текстовых объектов: от единичного текста до политематической коллекции текстов.

В рамках проводимого исследования рассматривались:

- максимально неоднородная – и по тематическим, и по стилевым характеристикам – коллекция новостей из русскоязычного сегмента веб-пространства;

- поэма Н.В.Гоголя «Мертвые души» (первый том).

На уровне выбора материала мы пытались максимизировать количество противопоставлений:

1) новостной *vs* художественный функциональный стиль,
2) коллекция *vs* одно произведение,
3) тематическая и стилевая неоднородность (новостей) *vs* однородность (поэмы Н.В. Гоголя).

Исследовалась зависимость локальной частоты встречаемости слов от глобальной с тремя значениями окна анализа (*K*=100, *K*=500 и *K*=5000).

Окна анализа подбирались эмпирически, их выбор был обусловлен желанием в качестве минимального окна выбрать тот диапазон, в который помещается средний абзац для поэмы или средний текст новостей (*K*=100), в качестве максимального окна – средняя глава поэмы или сегмент, в котором реализуется большинство новостных текстов, реализующих наиболее распространенную и актуальную новость (*K*=5000).

Цель исследовала состояла в том, чтобы на основании сопоставления частот встречаемости слов выделить основные единицы анализа для структур, описывающих коллекцию и/или текст. Для художественного произведения, скорее всего этой единицей будет сверхфразовое единство (СФЕ).

Формализовать критерии определения/выделения СФЕ в лингвистике текста, как правило, не удается. «Чистые СФЕ» встречаются крайне редко даже для текстов с максимальной однородностью тематики и стилевых характеристик.

Даже для самых однородных текстов наблюдается иерархия тем и отсутствие полной однородности стиля.
Противопоставление

текст *vs* коллекция-поток оказывается

динамическим, лишенным четких границ.

Семантической структурой называем структуру, характеризующую прежде всего стилевые характеристики.

Информационной структурой – структуру, характеризующую тематику, предметную область анализируемых текстов или коллекций. Для новостных (или научных) текстов эти структуры противопоставлены существенно выше, чем для художественных текстов [5].

5. *Ягунова Е.В., Пивоварова Л.М.* Экспериментально-вычислительные исследования художественной прозы Н.В. Гоголя. М., 2011

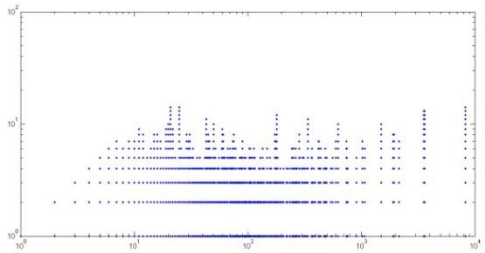
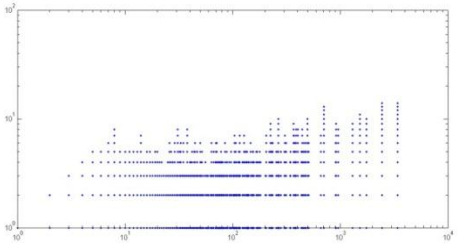
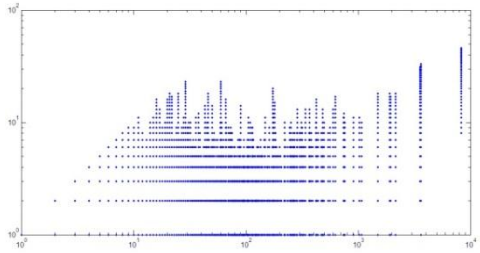
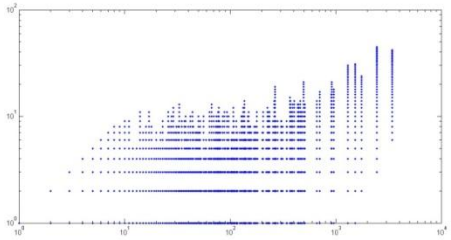
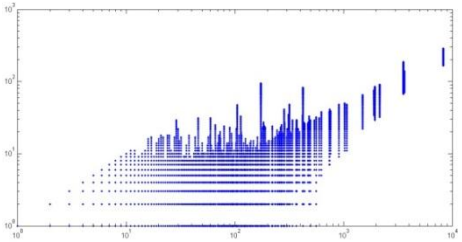
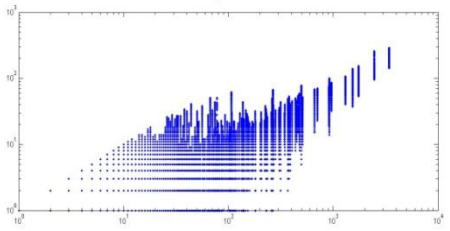
Ключевые слова с рассматриваемыми частотными характеристиками

Глобальная частота встречаемости	Ключевое слово с рассматриваемыми динамическими частотными характеристиками
128	ЧЕЛОВЕК
107	**НОЗДРЕВ**
73	**СОБАКЕВИЧ**
67	**МАНИЛОВ**
63	**ДУШИ**
52	СЕЛИФАН
54	**ЧИЧИКОВ**
43	**МЕРТВЫЕ**
38	ПРЕДСЕДАТЕЛЬ
33	ИВАН
29	**КАПИТАН**
26	**КОПЕЙКИН**
17	АНТОНОВИЧ

Ключевые слова, полученные в результате эксперимента с информантами

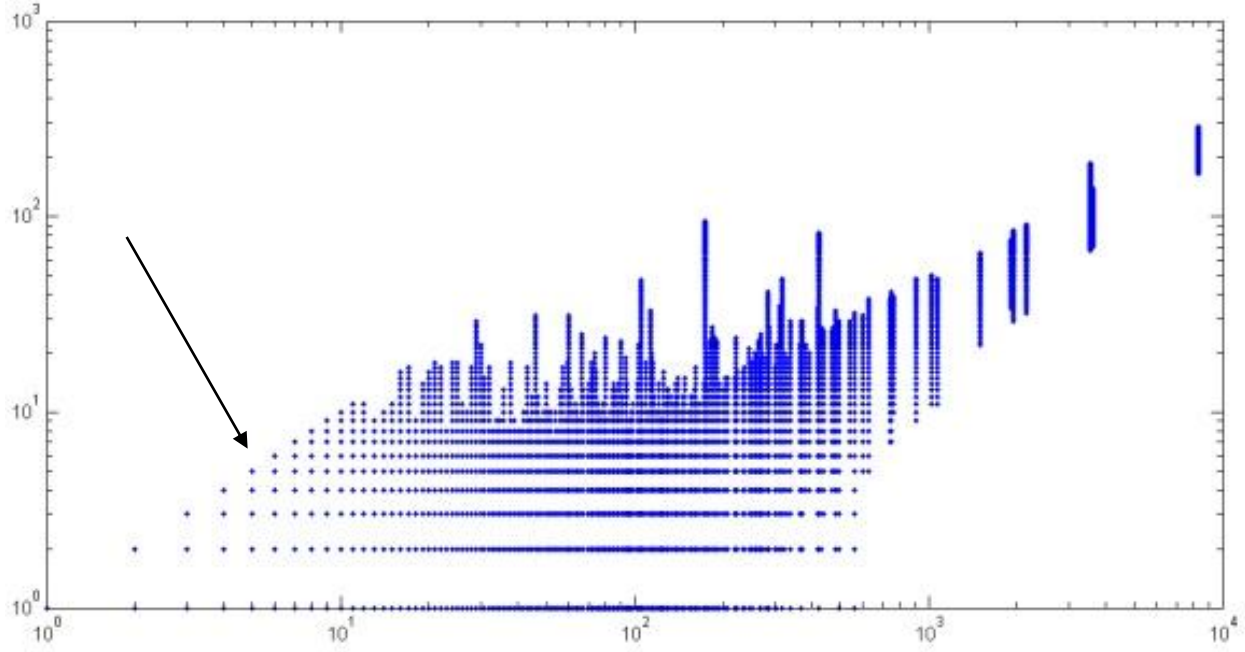
№ п/п	Ключевые слова	Вес
1	ПОМЕЩИК	10
2	БРИЧКА	8
3	ТРОЙКА	8
4	**ЧИЧИКОВ**	8
5	ДОРОГА	7
6	КОРОБОЧКА	7
7	ПЛЮШКИН	7
8	КУПЧАЯ	6
9	**МАНИЛОВ**	6
10	**СОБАКЕВИЧ**	6
11	**МЕРТВЫЕ ДУШИ**	6
12	ГУБЕРНАТОР	5
13	**НОЗДРЕВ**	5
14	КРЕПОСТНЫЕ	3
15	РОССИЯ	3

Графики зависимости локальной частоты от глобальной для различных *K*. Очевидно, при *K*→*N*, где *N* – общее число слов в анализируемом объекте, верхняя кромка графика будет стремиться к прямой (локальная частота станет совпадать с глобальной).

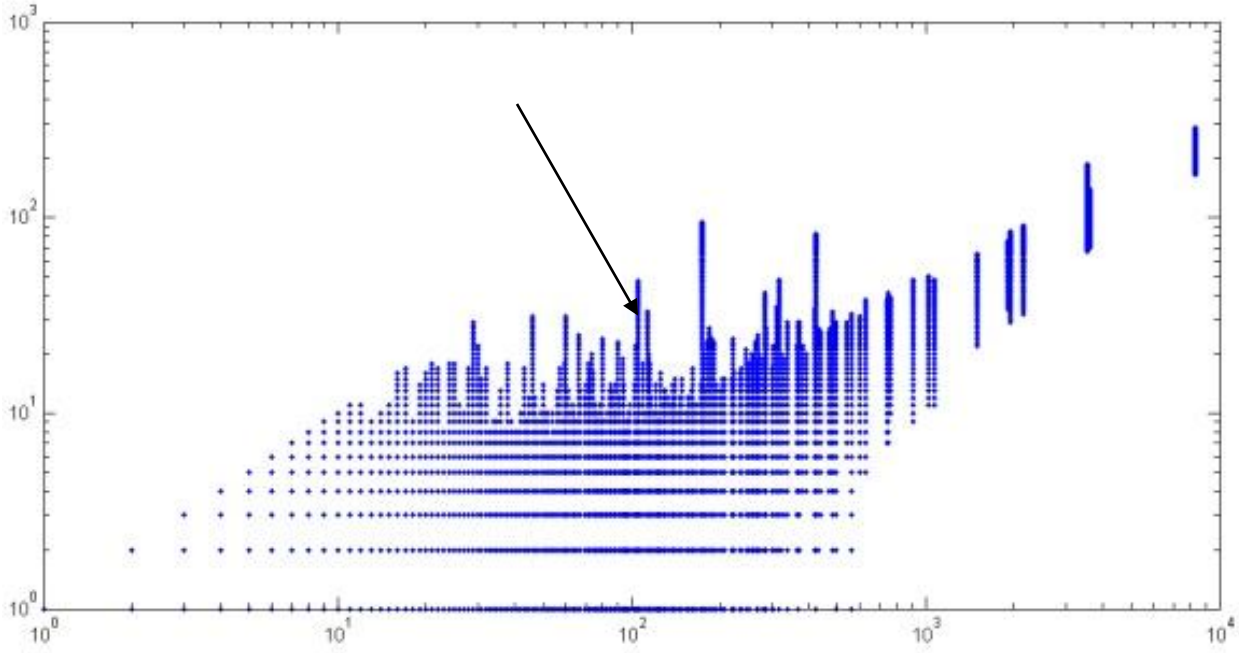
	Массив из веб-пространства:	«Мертвые души», том 1
K=100:		
K=500:		
K=5000:		

На каждом графике выделяется 4 области в соответствии со следующими параметрами:

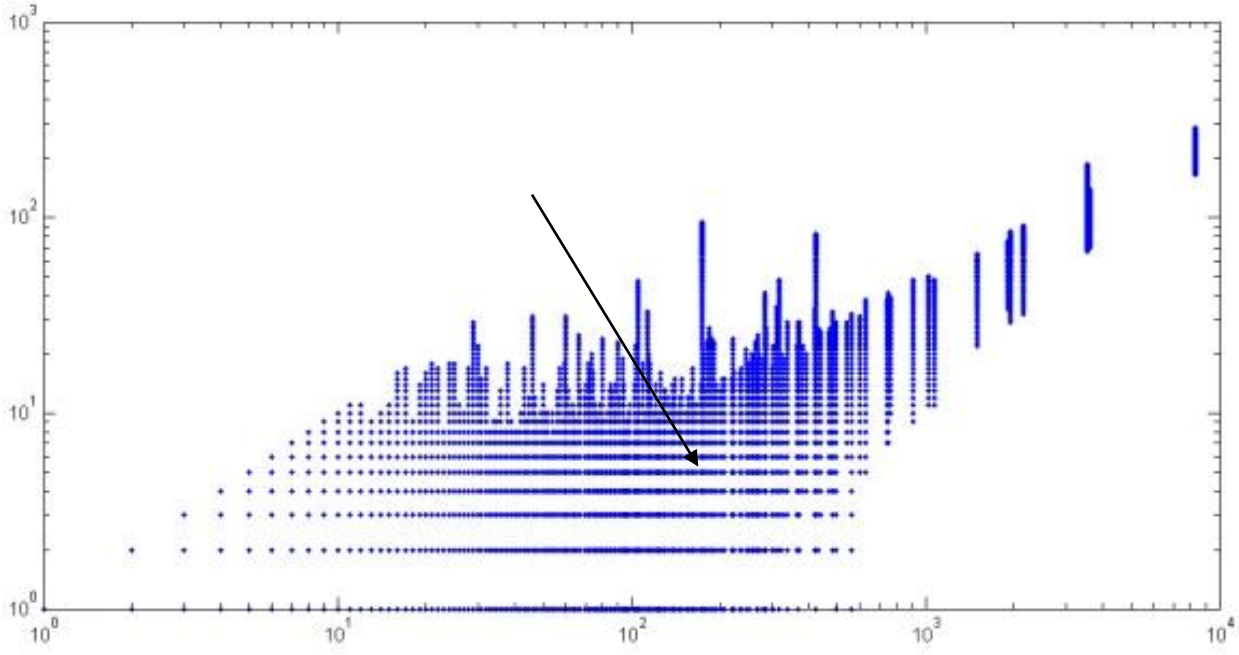
1. Глобальная и локальная частот малые. Таких слов очень много, их значение в тексте соответствует «хвосту» распределения Ципфа – это, прежде всего, редко используемые специфические слова, т.е. слова, характеризующие данный документ (сегмент потока) и встречающиеся более одного раза как глобально, так и локально. Кроме таких специфических слов в «область 1» попадают ошибки, которые достаточно легко отфильтровать.



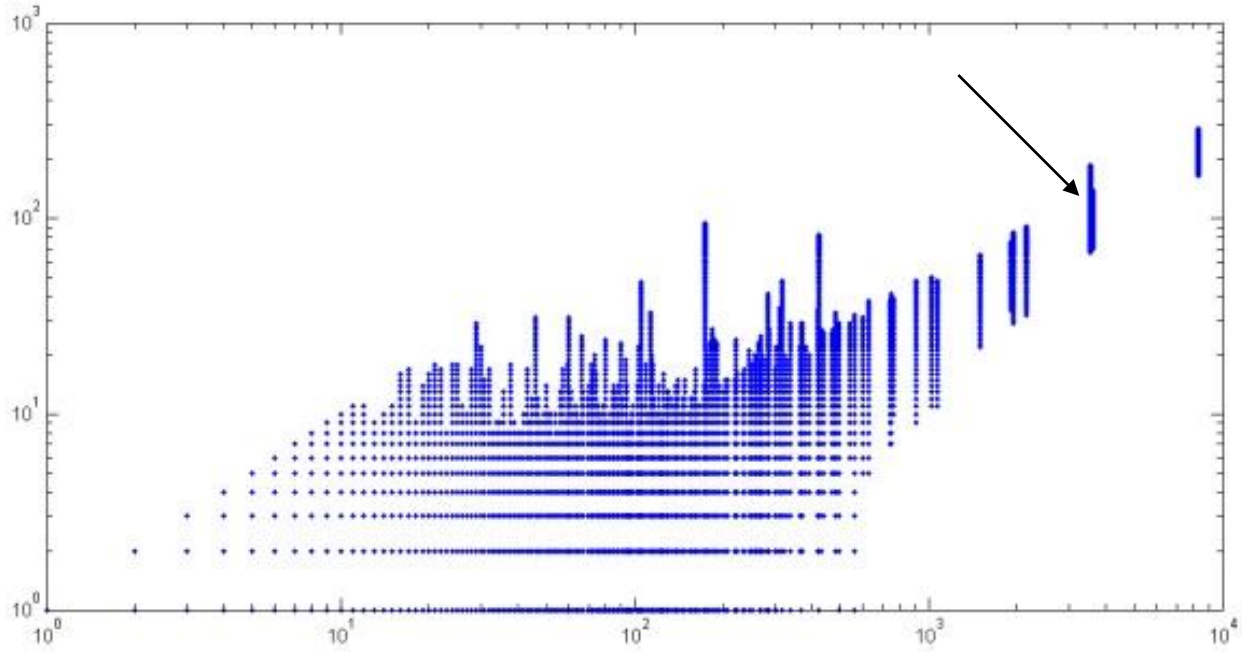
2. Глобальная частота относительно небольшая, а локальная – высокая. Этой области соответствуют слова, присущие новой теме, «всплеску» интереса к определенному факту в потоке новостей на сравнительно небольшом временном сегменте веб-пространства. Этой области соответствуют слова единичного текста, маркирующие СФЕ с наиболее четкими границами, например, появление действующего лица, локализованного в данном СФЕ (сегменте текста) и сопровождаемого «всплеском» внимания.



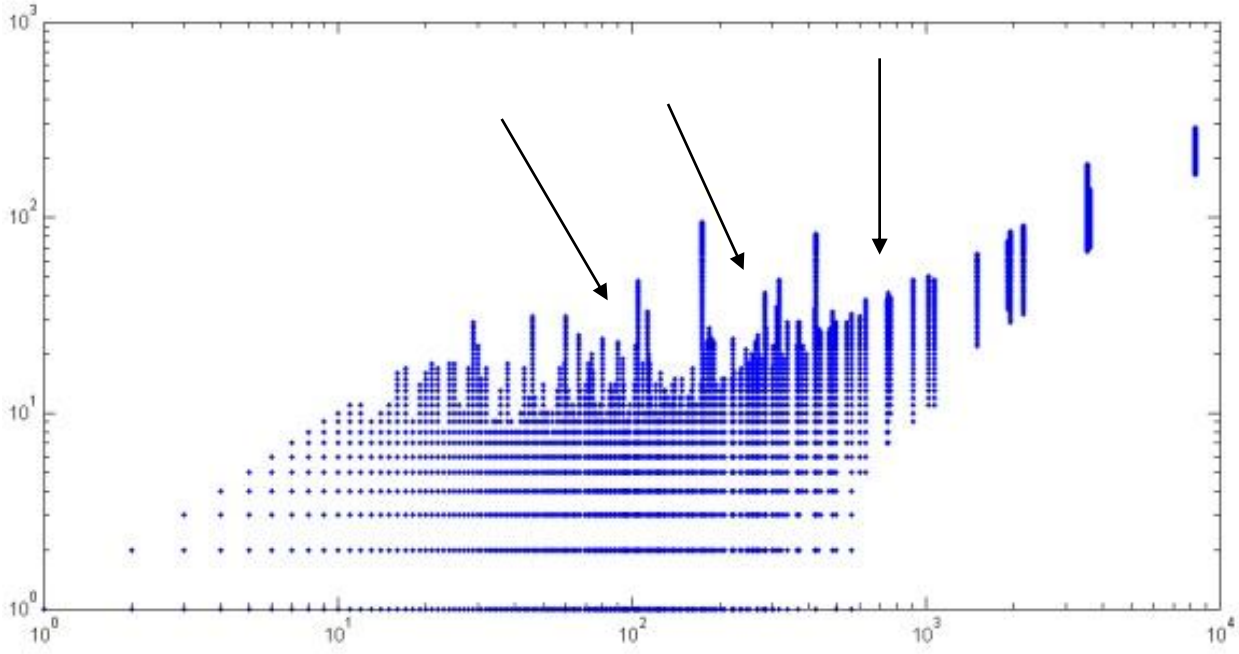
3. Глобальная частота высокая, а локальная – низкая. Этой области соответствуют слова относительно равномерно входящие в текст, по-видимому, определяющие его общую структуру: прежде всего, семантическую структуру, в которой задаются общие стилевые характеристики анализируемого объекта (текста и/или коллекции) и способ «упаковки» информации. Вероятно, это те слова, которые соответствуют скорее «семантической структуре» текста, в отличие от «информационной структуры», к которой по преимуществу относятся слова из п.2.



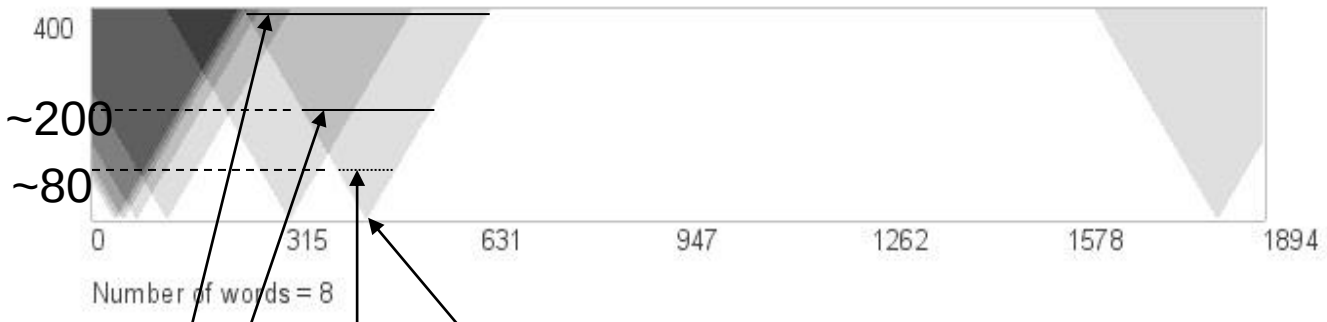
4. Глобальная и локальная частоты высокие. Чаще всего служебные слова, имеющие низкую «различительную силу» при поиске, такие слова обычно помещаются в список «стоп-слов».



Сосредоточимся на словах, у которых глобальная частота уже большая, а локальная скачет («гребешок»). Это наиболее информативный фрагмент (взаимодействие между областями и структурами).
Для поэмы «Мертвые души» практически все знаменательные слова являются теми ключевыми словами, которые явно маркируют СФЕ, сопровождаемые всплеском внимания на соответствующие реалии: человек, Ноздрев, Собакевич, Манилов, души, Чичикова, Селифан, мертвые, председатель, капитан, Копейкин. Назовем эти слова ключевыми, т.к. они совпадают с теми списками, которые выделяли информанты и/или с наибольшими значениями *TF IDF*.



Принцип построения спектограмм



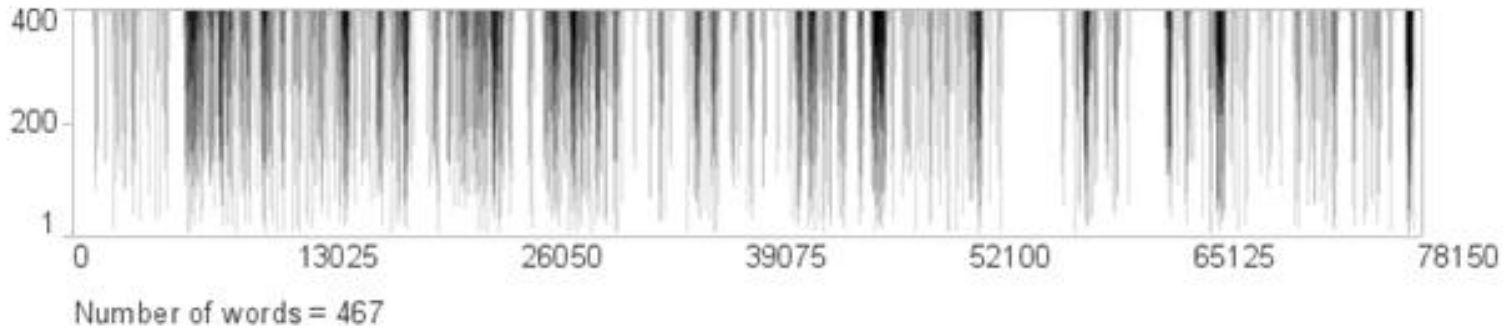
В этой позиции найдено слово

Окно наблюдения примерно 80 слов –
в нем пока слово только одно

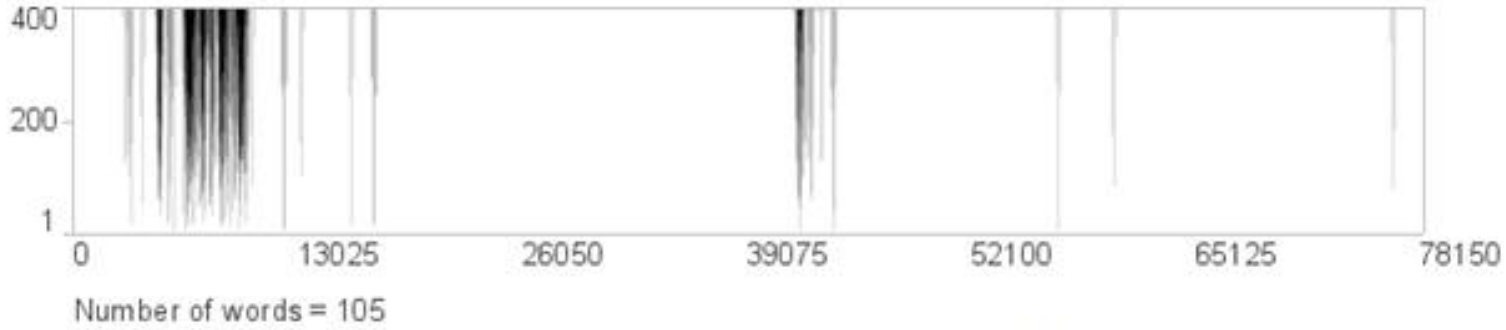
Окно наблюдения примерно 200 слов –
в нем найдено 2 слова

Окно наблюдения примерно 400 слов –
в нем найдено 4 слова – это видно по расцетке
наиболее темного участка

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 / - К.: Довіра. - С. 63-64.

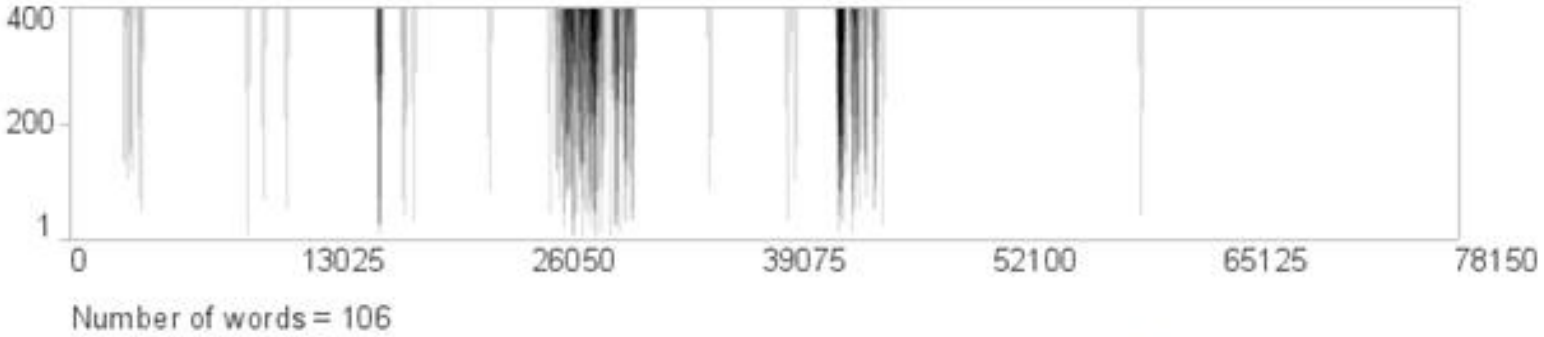


Спектрограмма для лексемы «Чичиков»

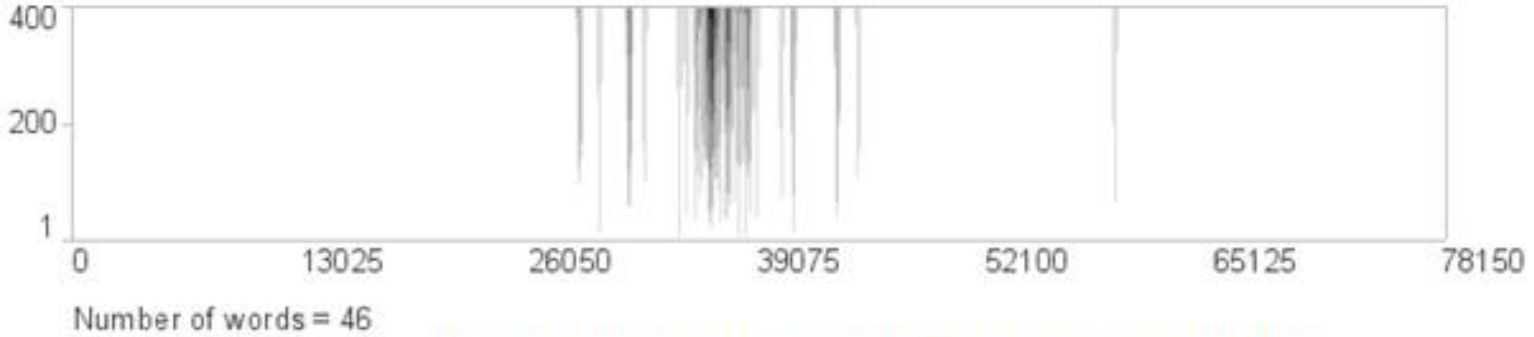


Спектрограмма для лексемы «Манилов»

Спектрограмма для лексемы «Ноздрев»



Спектрограмма для лексемы «Собакевич»



Спектрограмма для лексемы «Плюшкин»

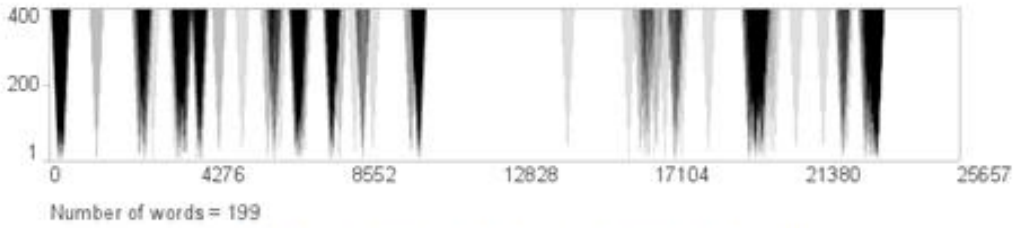
Спектрограмма для лексемы «Копейкин»

На материале новостной коллекции ключевые слова ведут себя еще более явным образом, их роль по сравнению с незнаменательной лексикой гораздо выше, чем для однородного единичного текста художественной литературы.

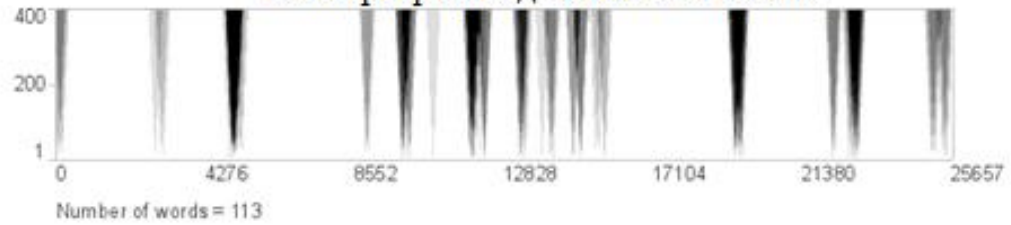
Проиллюстрируем это положение на примере локальных информационных всплесков начала декабря 2008 года: ОПЕК («Президент ОПЕК пригласил Россию вступить в картель»), РЖД («Из-за кризиса РЖД в ноябре сократила грузоперевозки на 20 процентов»), нефти («Распоряжение о строительстве нефтепровода в обход Белоруссии»); примеры государственный и университет иллюстрируют соединение двух словоформ в сложный термин (биграмму) («Не принимать абитуриентов по ЕГЭ разрешили 24 вузам»).

Ключевые слова из потока новостей, полученные в результате вычислительного эксперимента

Глобальная частота встречаемости	Ключевые слова с рассматриваемыми динамическими частотными характеристиками
370	ДОЛЛАР
340	ПРОЦЕНТ
286	США
175	РУБЛЬ
149	СУД
90	НАТО
66	НЕФТЬ
60	NOKIA
50	YAHOO
46	ОПЕК
29	РНК
28	РОНАЛДУ
22	РЖД
22	МЭРИ
21	ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
20	VISTA
16	DIXIS
15	FACEBOOK
11	SANYO



Спектрограмма для лексемы «банк»



Спектрограмма для лексемы «газ»

Спектрограмма для лексемы «доллар»

Спектрограмма для лексемы «нефть»

На данном примере видно, что слова «банк», «газ» и, с некоторой натяжкой, слово «доллар» маркируют СФЕ, в то время, как слово «нефть» не обладает такими свойствами.

Можно ли назвать сегменты новостного потока, выделенные благодаря локальным всплескам, аналогами СФЕ? Да, безусловно. Каждый из них описывает одну ситуацию, характеризуется максимальной тематической и стилевой однородностью. Более того, то, что выделяется по предлагаемой методике, как правило, хорошо локализовано, имеет явно выраженные временные и тематические границы.

В заключение подчеркнем, что современная лингвистика ориентирована на разнообразие лингвистических объектов: от традиционного объекта, эквивалентного единичному тексту, до коллекций и потоков новостей. И предлагаемый метод, ориентирован на исследование разных лингвистических объектов, когда единичный текст перетекает в поток текстов, а лингвистика текста смыкается с лингвистикой Интернета.

СПАСИБО ЗА ВНИМАНИЕ!

Ягунова Елена Викторовна,
iagounova.elena@gmail.com

Ландэ Дмитрий Владимирович,
dwlande@gmail.com