Динамические частотные характеристики как основа для структурного описания разнородных лингвистических объектов

© Е.В. Ягунова

С.-Петербургский гос. Университет

Санкт-Петербург
iagounova.elena@gmail.com

© Д.В. Ландэ
Институт проблем регистрации информации
НАН Украины
Киев
dwlande@gmail.com

Аннотация

В статье представлен подход к изучению динамических характеристик слов для описания разнородных динамических объектов: от отдельных текстов до потоков новостей. Четыре группы слов, выделяемых на основании динамических частотных характеристик, имеют четкую физическую и языковую природу. Они соответствуют разнородным по лингвистическим характеристикам объектам (с точки зрения структуры объекта и особенностей языка).

Введение

Изменившиеся условия существования человечества в условиях перехода к информационному обществу коренным образом перестроили процедуры анализа информации. Развитие поисковых технологий открыло ***новое*** поле деятельности для специалистов в области компьютерной лингвистики текста. Раньше основным и единственным объектом лингвистического исследования был ***текст*** (его анализ, понимание). Вместе с тем, объемы информации, содержащейся в информационных потоках, не могут быть восприняты и проанализированы отдельным человеком в силу его психофизиологических ограничений. Новый объект – ***информационный поток*** – требует использования новых технологий, которые выступают в качестве посредника при извлечении адресатом коммуницируемого смысла. Информационный поток в рамках данной работы рассматривается как совокупность сообщений, циркулирующих в информационном пространстве, в частности, его репрезентативной части - вебпространстве. Как правило, предметом научных исследований в настоящее время выступают информационные потоки, соответствующие определенным темам, которые на практике могут определяться тематическими запросами – фильтрами политематического информационного потока. В нашей же работе исследуется политематический информационный поток, который понимается, прежде всего, как ***множество текстов***, ***выступающих как некоторое единство***: адресатов интересует смысл, заключенный сразу в сотнях и тысячах текстов. В качестве «слепка» информационного потока в данной работе исследуется массив веб-публикаций из RUNeta, заиндексированный системой InfoStream (http://infostream.ua) в течение декабря 2008 г., объем которого составляет около 1200000 сообщений (~3500 источников).

Что такое текст? Даже структура единичного текста исследована более чем неполно. Прежде всего, текст – это основная единица, в которой содержится коммуницируемый смысл. В рамках традиционного лингвистического подхода основные характеристики текста определяют как:

- развернутость, или «последовательность знаковых единиц» (например, [1]);
- отдельнооформленность;
- связность и цельность.

Развернутость соотносится с вопросом о размерности и уровне иерархии такой единицы, как текст, структурными составляющими которого являются слова, синтагмы, фразы, сверхфразовые единства (забегая вперед – и далее). Выделяют «внешнюю» и «внутреннюю» (смысловую) связность. В основе связности и цельности текста – взаимосвязанность и взаимообусловленность его структурных составляющих. Связность реализуется как пространственная (контактно расположенные структурные составляющие), «логическая» и ассоциативная. В последних исследованиях все чаще разделяют когезию и когерентность (например, см. [2]. Когезия – связь элементов текста, при которых интерпретация одних элементов зависит от других [2]. Когерентность соотносима с прагматической стороной, она выводит нас за пределы текста в коммуникативную ситуацию и опирается на базу знаний адресата. Когерентность в наибольшей степени связана с реализаций (смысловых) ожиданий адресата. Однако часто невозможно четко разграничить эти два разных вида связности.

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

Природа связей может быть различного происхождения:
(1) связанной с лексической и семантической сочетаемостью/несочетаемостью,
(2) определяющейся правилами синтаксиса,
(3) соотносимой с информационной значимостью,
(4) задаваемой коммуникативной ситуацией вообще и задачей коммуникации в частности.

Применимы ли основные характеристики текста к такому лингвистическому объекту как информационный поток? По-видимому, ответ должен быть положительным. Наиболее проблемной характеристикой является отдельнооформленность, которая предполагает, с одной стороны, наличие сигналов начала и конца, а с другой – представление о фреймах: знании носителей языка о структуре текстов (текстовой и коммуникативной компетенции). Но проблемы с наличием сигналов начала и конца возникают гораздо раньше, т.е. уже на уровне сложного текста. А то, что сегменты информационного потока содержат коммуницируемый смысл (и при решении определенных задач коммуникации эти объекты становятся основными), полагаем, ни у кого не вызывает сомнений. Очевидно и наличие большого количества лингвистических технологий, позволяющих решать эти задачи.

В рамках исследования частотных характеристик текста с целью сопоставления этих характеристик введем два определения:

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

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

Очевидно, что при исследовании информационного потока, количество записей будет изменяться в течение времени, т.е. такая характеристика, как глобальная частота встречаемости может ***динамически*** изменяться во времени. В данной работе она фиксируется нами для окна наблюдения, равного объему исследуемого фрагмента информационного потока или исследуемого текста литературного произведения. Локальная частота встречаемости же зависит от относительно небольшого скользящего окна наблюдения и может ***динамически*** изменяться в пределах всего текстового массива или единичного исследуемого текста литературного произведения.

Конечно, локальная частота встречаемости может быть разной для различных фрагментов текста, а также зависеть от перестановки сообщений в информационном потоке, однако, как показывает практика, при достаточно больших объемах наблюдаемых данных, характер их распределения, в частности, для массива веб-публикаций зависит лишь от величины окна наблюдения. Кроме того, инвариантность от перестановок отдельных сообщений обеспечивается нами в дальнейшем выбором величины окон наблюдений, которые по порядку соответствуют размеру среднего сообщения из информационного потока.

Данные характеристики являются ***динамическими***, так как в них учитывается динамическая картина взаимодействия частот встречаемости в пространстве анализируемого объекта (от коллекции до текста).

Цель исследования состояла в том, чтобы на основании сопоставления частот встречаемости слов – глобальной и локальной – выделить основные единицы анализа для структур, описывающих коллекцию и/или текст. Для описания выделяемых единиц попробуем использовать те наработки, которые может предоставить лингвистика текста, т.е. дополнительной целью работы является соединение методов и задач информационных технологий мониторинга новостей и лингвистических подходов в рамках нашего расширенного понимания объекта, методов и задач лингвистики текста.

Что могут представлять основные единицы анализа для структур, описывающих коллекцию и/или текст? Для художественного повествования, скорее всего такой единицей будет сверхфразовое единство (СФЕ) [3]. Это единица, которая наименее формализована в традиционной лингвистике текста, прежде всего, это относится к критериям определения/выделения подобных единиц. Чаще всего, когда описывают СФЕ, речь идет об единстве ситуации (событии): единстве действующих лиц, места, времени, а иногда и способе действия (или о некотором сходном составе). Таким образом наблюдается аналогия между описанием лингвистики текста и описанием в рамках информационного подхода, когда существенное внимание уделяется именно этим типам именованных сущностей.

Для новостного потока, вероятно, будут выделяться единицы, являющиеся некоторыми аналогами СФЕ (наиболее четкие из возможных аналогов СФЕ). Эти единицы предположительно состоят более чем из одного документа/ текста (значимая новость не должна быть представлена лишь одним документом). Это будет похоже на сегмент потока, состоящего из документов с максимальными локальными актуальностью и новизной. Вероятно, такой сегмент имеет сравнительно четкие временные границы начала и конца фрагмента. Таким образом, интересующая нас структурная единица в новостном потоке представляет собой единицу, размерность которой варьирует от новостного текста до кластера текстов, относящихся к одному временному сегменту и одной тематике (одна или ряд сходных ситуаций, объединяемых на основе наименований персон, организаций, географических наименований, времени и собственно наименования ситуации (события).

В чистом виде такие единицы встречаются крайне редко даже для текстов с максимальной однородностью тематики и стилевых характеристик. Почему? Потому что даже для самых однородных текстов наблюдается иерархия тем (тем и подтем) и отсутствие полной однородности стиля. В случае новостных текстов – потому что одна и та же ситуация может быть освещена по-разному даже в текстах информационных лент (уж не говоря о

тематических текстах с элементами аналитики), один и тот же текст может содержать информацию о нескольких ситуациях (событиях). В этом смысле противопоставление текст *vs* цикл *vs* коллекция-поток оказывается динамическим, лишенным четких границ.

В теории информационного поиска признано ранжирование весов слов по классическому критерию Солтона ***TF IDF*** [4], где *TF* (*Term Frequency*) – это частота встречаемости слова в пределах выбранного документа, а *IDF* (*Inverse Document Frequency*) – функция (чаще всего логарифм) от величины, обратной количеству документов, в которых встретилось данное слово. Наш подход идеологически близок к TF, можно считать, что локальная частота – это аналог *TF* (в этом случае окно наблюдения – аналог документа), а глобальная частота встречаемости соответствует *DF* (*Document Frequency*). При этом появляется возможность анализировать не только массивы документов, как это реализовано с помощью *TF IDF*, но и цельные тексты больших объемов (ср. [5]).

Следует отметить, что если в задачах информационного поиска достаточно часто исследуется поведение *TF IDF* (или некоторых близких по смыслу функций), в то время, как в рамках данной работы фактически исследуется взаимная зависимость двух сомножителей *TF* и *DF*.

В наших исследованиях ***приоритетное значение имеет весь текстовый массив*** (в отличие от каждого отдельного документа), значения глобальной частоты встречаемости не понижается путем логарифмирования как в *TF IDF*. Кроме того, критерий соотношения локальной и глобальной частоты встречаемости слов может применяться не только к слову из определенного фрагмента текста, но и позволяет видеть общую частотную картину, связанную с выбранным словом, оценивать его значение для всего текстового массива.

В [6] 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*) [7]. Однако нам не известны такого рода исследования, выходящие за рамки решения узко формулируемых задач мониторинга новостных потоков, например, на уровне слов, фрагментов текста, текста и т.д.

На наш взгляд предлагаемый подход позволяет анализировать структуры самых разных текстовых объектов: от единичного текста до политематической коллекции текстов, рассматриваемой как сегмент информационного потока.

Кроме того, предлагаемый подход позволит приблизиться к формализованной оценке такой составляющей, как СФЕ.

1 Материал и методика

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).

2 Результаты

Введем еще определения, с которыми мы отчасти будем соотносить свои результаты. ***Семантической структурой*** называем структуру, характеризующую прежде всего стилевые характеристики, ***информационной структурой*** – характеризующую тематику (предметную область) анализируемых текстов или коллекций. Для новостных (или научных) текстов эти структуры противопоставлены существенно выше, чем для художественных текстов [8].

На рис. 1 представлены графики зависимости локальной частоты от глобальной для различных окон анализа (K). Очевидно, при приближении значения K к общему числу N слов в анализируемом объекте (тексте и/или коллекции), верхняя кромка графика будет приближаться к прямой (локальная частота станет совпадать с глобальной).

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 (далее – веба). Этой области соответствуют слова единичного текста, маркирующие интересующие нас структурные единицы (сегменты текста) с наиболее четкими границами, например, появление действующего лица (и/или объекта), локализованного в данной единице (сегменте текста) и сопровождаемого «всплеском» внимания. Рассматриваемые слова почти наверняка относятся к информационной структуре. Мы абстрагируемся от проблем повторных номинаций, что позволительно именно на таких сегментах, т.к. высокий уровень внимания «заставляет» авторов текстов многократно повторять основную номинацию.

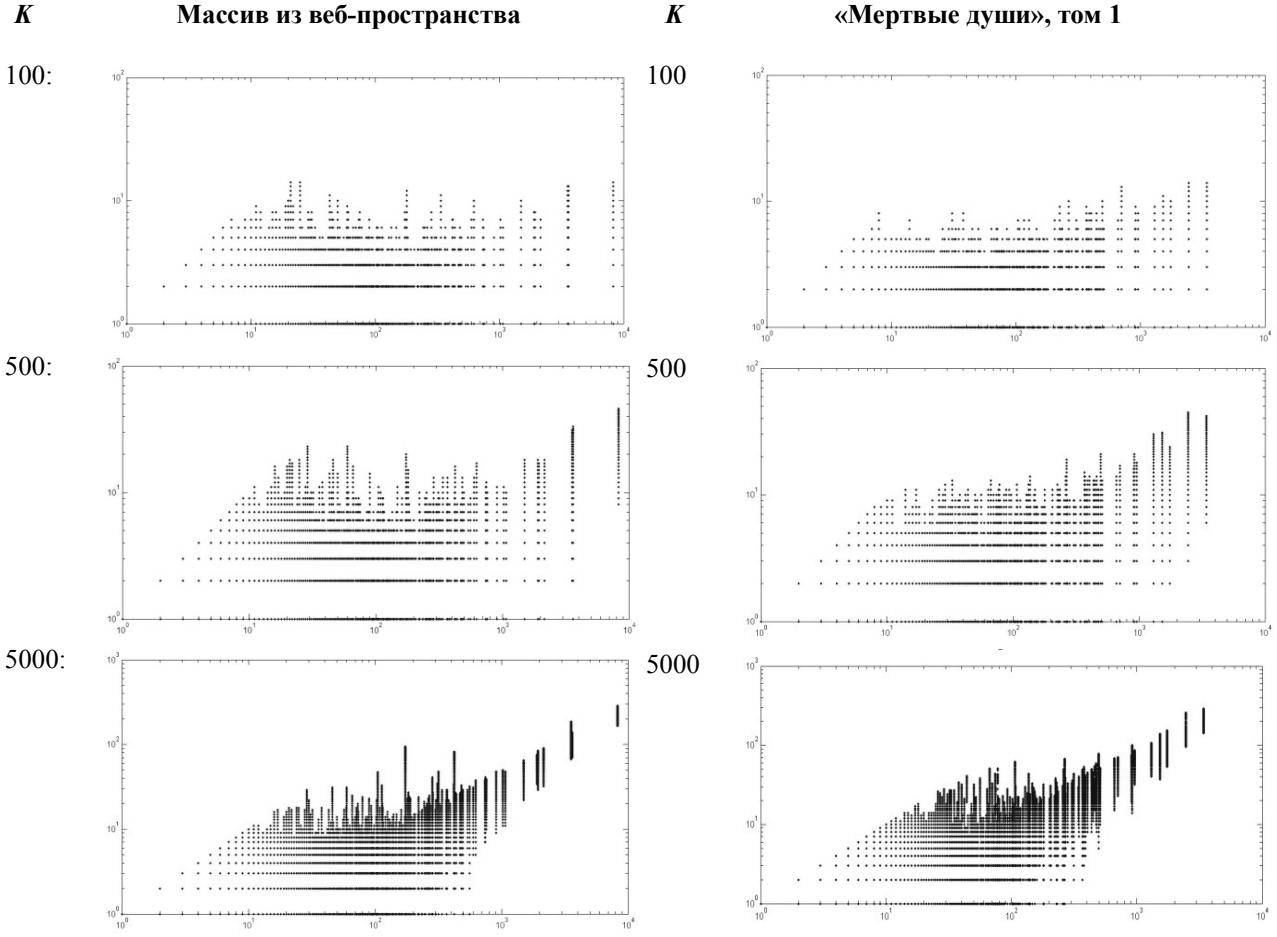

Рис. 1 *Зависимость локальной частоты встречаемости (вертикальная ось) от глобальной (горизонтальная ось) в двойной логарифмической шкале*

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. В данном случае – те характеристики, которые свойственны большинству новостных источников (из веба), или те, которые свойственны поэме «Мертвые души». Вероятно, это те слова, которые соответствуют скорее *семантической* структуре» текста, в отличие от *информационной* структуры, (к информационной структуре по преимуществу относятся слова из п. 2) (о разделении этих структур см. подробнее [9]).

4. *Глобальная и локальная частоты высокие*. В эту область чаще всего попадают служебные слова, имеющие низкую «различительную силу» при поиске, такие слова обычно помещаются в список «стоп-слов».

3 Обсуждение результатов

В данной статье мы сосредоточились на словах, у которых ***глобальная частота уже большая, а локальная скачет*** (см. «гребешок» на рис. 1). Это промежуточный и наиболее информативный для нас

фрагмент (взаимодействие между областями и структурами).

Для поэмы «Мертвые души» практически все знаменательные слова (имена) с такими частотными характеристиками являются теми ключевыми словами, которые явно маркируют СФЕ. СФЕ с этими словами сопровождаются всплеском внимания на соответствующие реалии при развертывании текста. Эти слова, упорядоченные по значению глобальной частоты, приведены в табл. 1, у них глобальная частота меньше 250, а локальная максимальная больше 10.

Табл. 1 *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

Эти слова можно считать ключевыми, так как все они соотносятся с теми наборами ключевых слов, которым соответствовали наибольшие значения *TF IDF,* более трети из них – с наборами слов, которые выделяли информанты (ср. [9]).

Так, для получения набора ключевых слов был проведен эксперимент (21 информант) с традиционной методикой и стандартной инструкцией [10]: «Вспомните «Мертвые души» Н.В. Гоголя. Подумайте над их содержанием. Выпишите 10-15 слов, наиболее важных для их содержания». Единственное отличие от традиционного варианта заключалось в том, что информантам предлагалось вспомнить тексты, т.е. оценивалось остаточное знание текста. В качестве информантов выступали, главным образом, 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. В табл. 2 приведены результаты эксперимента; вес слова (или словосочетания) приводится в абсолютных числах (указывается число информантов, записавших в анкете данное слово). Как и в вычислительном эксперименте, анализировались слова, которые могли объединяться в сложные номинации (*мертвые души*) только на основании анализа анкет.

Табл. 2 *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

Для того, чтобы выделить ключевые слова по мере *TF IDF* важно правильно определить **контекст**, а именно контрастивную коллекцию.

В качестве двух вариантов контрастивной коллекции рассматривались:

- «Гоголь+Чехов» – коллекция текстов Н.В. Гоголя (кроме «Мертвых Душ») и коллекция текстов А.П.Чехова (сборники «Человек в футляре», «Рассказы 1887 год», «Рассказы. Повести. 1888-1891», «Рассказы. Повести. 1892-1894», «Рассказы. Повести. 1894-1897»;
- «Гоголь» – коллекция текстов Н.В. Гоголя (кроме «Мертвых Душ»).

Выбор контекста определяется требованием максимальной однородности и опирался как на интуицию исследователя, так и на данные предварительного статистического анализа.

В табл. 3 приведены ключевые слова, выделенные с использованием меры важности *TF IDF*, слова упорядочены по убыванию значения этой меры. Пороговое значение определялось на основании графического изображения распределения значений меры; значение подбиралось так, чтобы набор был представительным для последующей интерпретации, а порог находился перед так называемым плато (последовательностью с близкими значениями меры).

Табл. 3 *Ключевые слова, полученные в результате вычислительного эксперимента*

Ключевые слова в контексте «Гоголь+Чехов»	Ключевые слова в контексте «Гоголь»
ЧИЧИКОВ	**ЧИЧИКОВ**
НОЗДРЕВ	**НОЗДРЕВ**
МАНИЛОВ	**МАНИЛОВ**
СЕЛИФАН	**СЕЛИФАН**
СОБАКЕВИЧ	ПАВЕЛ
КОСТАНЖОГЛО	**СОБАКЕВИЧ**
ЧЕЛОВЕК	**ПРЕДСЕДАТЕЛЬ**
ПЛЮШКИН	КОСТАНЖОГЛО
ПЛАТОН	ГЕРОЙ
ХЛОБУЕВ	ПЕТРУШКА
СЛОВО	ПЛЮШКИН
РУКА	ПЛАТОН
КОПЕЙКИН	ХЛОБУЕВ
МУРАЗ	БРИЧКА
АНТОНОВИЧ	ИМЕНИЕ
ПЕТРУШКА	ИВАНОВИЧ
БРИЧКА	**КОПЕЙКИН**
ПЛАТОНОВ	**ЧЕЛОВЕК**
ЛИЦО	ПОЛИЦЕЙМЕЙСТЕР
СОБАКЕВИЧА	РУКОПИС
КУПЧАЯ	ПОМЕЩИК
ПАВЕЛ	ПРОКУРОР
ГОРОД	ПРЕВОСХОДИТЕЛЬСТВО
СТОРОНА	КРЕСТЬЯНИН
ГЛАЗ	БАРИН
КОШКАРЕВ	ГУБЕРНАТОР
МЕСТО	КОНСТАНТИН
АССИГНАЦИЯ	СИЯТЕЛЬСТВО
ГЕРОЙ	АНДРЕЙ
ДУШИ	АФАНАСИЙ
ДАМА	МУРАЗ
ГОЛОВА	ДУШИ
ЛЕНИЦЫН	**АНТОНОВИЧ**
ПОЭМА	ХОЗЯЙКА
ЧУБАРЫЙ	ДЕРЕВНЯ
ДУМАТЬ	ГЕНЕРАЛ
ИВАНОВИЧ	ГОСТЬ
ЖИЗНЬ	БАТЮШКА
БОГ	ХОЗЯИН
ДОМ	КРЕПОСТЬ
БАРИН	ХОЗЯЙСТВО
ПОЛИЦЕЙМЕЙСТЕР	КНЯЗЬ
ПРЕДСЕДАТЕЛЬ	**МЕРТВЫЕ**

Следует отметить, что не все выделенные информантами ключевые слова попали в искомый список слов, у которых ***глобальная частота уже большая, а локальная скачет*** (см. «гребешок» на рис. 1). Кроме того, не все ключевые слова, характеризующиеся максимальными значениями *TF IDF* (см. табл. 2), попали в этот список. Эти факты, по-видимому, объясняются тем, что это список тех слов, которые явно маркируют определенные СФЕ, но не обязательно весь текст. Это список тех слов, которые сопровождаются всплеском внимания на соответствующие реалии в процессе развертывания текста.

Приведем несколько примеров из визуализации (распределения в тексте разных действующих лиц) с помощью сервиса [11] (рис. 2-7), доступного по адресу http://ling.infostream.ua/jag/. В рамках этого сервиса обеспечивается визуализация плотности встречаемости слова в тексте в зависимости от ширины окна наблюдения. В приведенных спектрограммах 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». Максимальное окно наблюдения в приведенных случаях составляет 400 словоупотреблений (с/у).

На рис. 2 представлена спектрограмма, отражающая распределение наименования главного действующего лица: лексема «Чичиков» в тексте встречается 467 раз. На рис. 3 представлена спектрограмма для лексемы «Манилов» (105 словоупотреблений в тексте (с/у)), на рис. 4 – для лексемы «Ноздрев» (143 с/у), на рис. 5 – для лексемы «Собакевич» (106 с/у), на рис. 6 – для лексемы «Плюшкин» (46 с/у), на рис. 7 – для лексемы «Копейкин» (32 с/у).

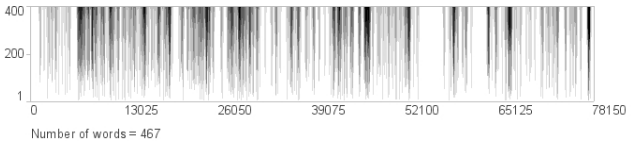

Рис. 2 *Спектрограмма для лексемы «Чичиков»*

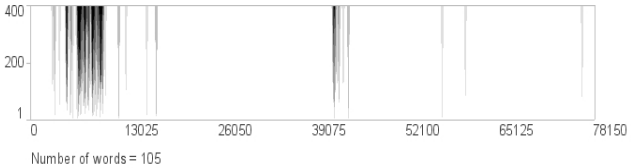

Рис. 3 *Спектрограмма для лексемы «Манилов»*

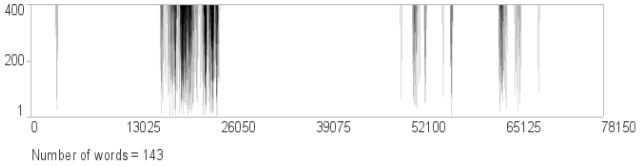

Рис. 4 *Спектрограмма для лексемы «Ноздрев»*

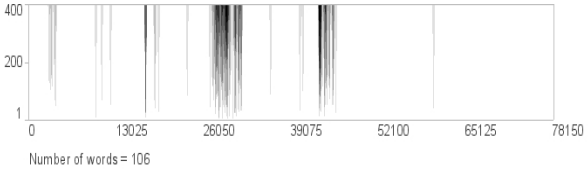

Рис. 5 *Спектрограмма для лексемы «Собакевич»*

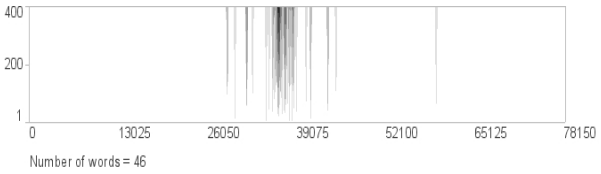

Рис. 6 *Спектрограмма для лексемы «Плюшкин»*

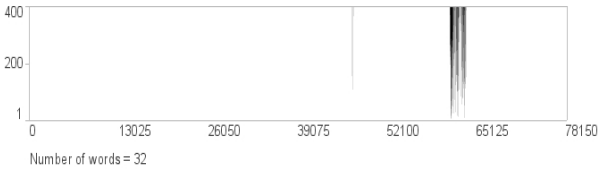

Рис. 7 *Спектрограмма для лексемы «Копейкин»*

Ключевые слова Чичиков, Манилов, Ноздрев, Собакевич, Копейкин являются теми ключевыми словами, которые явно маркируют интересующие нас структурные единицы текста (см. табл. 1). В отличие от них лексема «Плюшкин» сосредоточено на одном, но очень расплывчатом фрагменте, оно не может маркировать и привлекать внимание, т.е. служить для сегментации и идентификации соответствующей единицы.

Коммуникативные, модальные и некоторые другие классы глаголов (например, глаголы воображения), дискурсивные слова маркируют смену коммуникативных (как части семантических) структур: модальности, тональности, адресности и т.д. Эти слова приведены в таблице 4, они упорядочены по значению глобальной частоты.

Для «Мертвых душ» – среди слов, у которых глобальная частота уже большая, а локальная скачет – наиболее характерны глаголы «*говорить*», «*представить*», «*понимаете*» (выделено п/ж шрифтом в табл. 4). Для глагола «*представить*» самые актуальные конструкции – «*можете себе представить*» или «*можете представить себе*». Глагол «*понимаете*» (часто вместе с частицей «*ли*») несет чисто коммуникативную (и стилистическую, конечно) нагрузку (напр., в составе конструкций «*Понимаете ли?*», «*Понимаете?*», *«.., понимаете,...*»). Как составные элементы этих конструкциями «*можете, ли, себя*» находятся среди рассматриваемого списка слов, в табл. 4 они выделены полужирным курсивом. В списке слов присутствуют междометия («*О*» и «*Ах*» с разными знаками препинания), личные местоимения, частицы, наречия.

Таким образом выделяются структурные единицы, характеризуемые резкой сменой коммуникативных или модальных свойств: например, отстраненное повествование от третьего лица сменяется обращением к адресату (или диалогом). Разные виды повествователя и адресата в нарративе, речевой и нарративный режимы и т.д. в художественном повествовании уже неплохо изучены в лингвистической теории (см., например, Падучева 2010).

Табл. 4 *Ключевые слова из «Мертвых душ»*

Глобальная частота встречаемости	Ключевые слова с рассматриваемыми динамическими частотными характеристиками
230	***ЛИ***
222	ЕМУ
192	НЕТ
189	БЫЛИ
179	ВЫ
162	О
141	*СЕБЕ*
139	ОЧЕНЬ
137	НУ
129	ВЕДЬ
116	МНЕ
101	МЕНЯ
97	ТЕБЯ
94	ОНА
91	ТЕБЕ
83	ТАМ
58	**ГОВОРИТ**
42	АХ
29	***МОЖЕТЕ***
28	**ПОНИМАЕТЕ**
23	**ПРЕДСТАВИТЬ**

Предлагаемый подход, основанный на использовании динамических частотных характеристик, позволяет формализовать возможности описания семантической структуры текста в тесном взаимодействии с информационной структурой. Выйти на формализованный способ описания связности текста на уровне интересующих нас структурных единиц (тем самым приблизиться и к пониманию природы сегментации на СФЕ).

На материале новостной коллекции ключевые слова, у которых глобальная частота уже большая, а локальная скачет**,** ведут себя еще более явным образом, их доля по сравнению с незнаменательной лексикой гораздо выше, чем для однородного единичного текста художественной литературы. В табл. 4 приведен список анализируемых слов (у них, как и в рассматриваемом ранее случае, глобальная частота меньше 250, а локальная больше 10).

Проиллюстрируем это положение на примере локальных информационных всплесков начала декабря 2008 года (приводится название одного из документов): *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»), DIXIS («Судебные приставы арестовали имущество Dixis»). Yahoo («Microsoft подтвердил заинтересованность в покупке поиска Yahoo!»), Facebook

(«Сетевой червь атаковал компьютеры пользователей Facebook») и т.д. Примеры *государственный* и *университет* иллюстрируют соединение двух словоформ в неоднословную номинацию («Не принимать абитуриентов по ЕГЭ разрешили 24 вузам»).

Табл. 5 *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

Локальные максимумы на графиках (рис. 1), соотносимые с коммуникативными и модальными характеристиками коллекции (сегмента русскоязычного новостного веба), проявляются, например, в резком локальном всплеске определенной дискурсивной и/или местоименной лексики, особенно личных местоимений типа «*мы*» (*глобальная частота встречаемости* – 174, максимальная локальная – 19), «*я*» (*глобальная частота встречаимости* – 105, максимальная локальная – 11]). Такого рода лексических единиц выделяется гораздо меньше, но они оказывают не менее яркое влияние на то, что обычно называется дискурсом (дискурсивными практиками), в данном случае это важные локальные всплески, характеризующие новостной дискурс конца 2008 года.

На рис. 8-11 представлены спектрограммы, отражающие распределения слов «банк», «газ», «доллар», «нефть» в небольшом массиве новостных сообщений за указанный период из одного источника (elvisti.com – «Обзор основных событий дня») по тематикам экономика и энергетика.

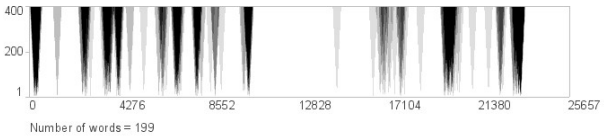

Рис. 8. *Спектрограмма для лексемы «банк»*

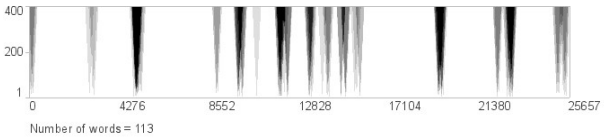

Рис. 9. *Спектрограмма для лексемы «газ»*

Рис. 10. *Спектрограмма для лексемы «доллар»*

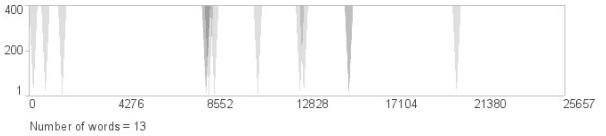

Рис. 11. *Спектрограмма для лексемы «нефть»*

На примере этих спектрограмм рассматривается степень равномерности / выделенности лексемы в небольшом новостном массиве одного источника для данного временного периода – видно, что лексемы «банк», «газ» маркируют интересующие нас структурные единицы этого новостного массива. Лексема «доллар» с некоторой натяжкой может маркировать определенные сегменты этого массива, хотя и с небольшим весом. Лексема «нефть» явно не обладает такими свойствами.

Заключение

На основании сопоставления частот встречаемости слов – глобальной и локальной – можно выделить основные структурные единицы, позволяющие описывать как новостную коллекцию, так и единичный текст литературного произведения. Для описания выделяемых единиц разумно использовать методы и подходы лингвистики текста, т.е. оптимальным для их описания является соединение методов информационных технологий и лингвистики текста. Каждая из выделяемых единиц 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. Поэтому сегменты новостного потока, выделенные благодаря локальным всплескам, можно назвать аналогами СФЕ?

Четыре группы слов, выделяемых на основании динамических частотных характеристик, имеют четкую физическую и языковую природу. Они характеризуют объекты (разнородные по лингвистическим характеристикам) с точки зрения структуры объекта и особенностей конкретного рассмат-

риваемого (под)языка (языка текста или рассматриваемой коллекции).

Два класса – класс слов с малыми глобальной и локальной частотами и класс с высокой глобальной и локальной частотами – соотносятся с распределением Ципфа (явным или метафорически понимаемым (если объем объекта не позволяет строить подобные распределения)).

Класс слов, у которых глобальная частота относительно небольшая, а локальная – высокая, маркирует моменты всплеска интереса (и сами объекты, вызывающие этот интерес). Речь идет, например, о «всплеске» интереса к определенному факту в потоке новостей (на сегменте веба (СФЕ)) или всплеск интереса к действующему лицу (и/или объекту) в пределах рассматриваемой структурной единицы текста. Эти слова выступают в качестве фигур на фоне остальных единиц сегмента (в терминах гештальт-психологии) и почти наверняка относятся к информационной структуре. С другой стороны, эти слова могут выделять и собирать вокруг себя соответствующие единицы (СФЕ), тем самым опосредованно сегментируя поток на такие структурные составляющие как СФЕ.

Класс слов, у которых глобальная частота высокая, а локальная – низкая, относительно равномерно распределены в тексте и определяют, прежде всего, семантическую структуру, в которой задаются общие стилевые характеристики анализируемого объекта (текста и/или коллекции) и способ «упаковки» информации. Это те слова, которые соответствуют скорее семантической структуре» текста, в отличие от информационной структуры.

В статье мы сосредоточились на словах, у которых глобальная частота уже большая, а локальная скачет. Это промежуточный класс (между третьим и четвертым) и наиболее информативный для нас фрагмент, т.к. именно на нем реализуется взаимодействие между информационной и семантической структурами. Когда мы рассматривали класс слов, у которых глобальная частота относительно небольшая, а локальная – высокая, мы понимаем, что в сегментации потока эти слова могут участвовать лишь опосредованно (опираясь на них, мы узнаем количество единиц, но не границы между ними). Как только мы переходим к промежуточному классу, в фокус внимания попадают как ключевые слова (принадлежащие информационной структуре), так и слова, образующие структуру и стиль (тем самым принадлежащие семантической структуре). Первые слова (ключевые) выделяют единицы: обозначают основные объекты, являющиеся фигурами в локальной структуре объекта. Слова, образующие структуру и стиль, по-видимому, часто маркируют границы, обозначая изменение коммуникативной стратегии (или нарративного режима в художественном тексте).

В заключение еще раз подчеркнем, что современная лингвистика должна быть ориентирована на разнообразие лингвистических объектов: от традиционного объекта, равного единичному тексту, до коллекций и потоков новостей (ср. [12]). И предлагаемый метод, ориентирован на исследование ***различных*** лингвистических объектов, когда единичный текст перетекает в поток текстов, а лингвистика текста смыкается с лингвистикой Интернета.

Литература

[1] *Николаева Т.М.* Краткий словарь терминов лингвистики текста // Новое в зарубежной лингвистике. Вып. VIII. – М., 1978.

[2] *Кронгауз М. А.* Семантика. – М.: РГГУ, 2001 г. 399 с.

[3] *Солганик Г. Я.* Синтаксическая стилистика. Сложное синтаксическое целое. – 2-е изд., испр. и доп. – М.: Высш. шк., 1991. – 182 с.

[4] *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval. Information Processing and Management, 1988. – № 24(5). – P. 513-523.

[5] *Ягунова Е.В.* Ключевые слова в исследовании текстов Н.В. Гоголя // Проблемы социо- и психолингвистики. Вып. 15: Пермская социолингвистическая школа: идеи трех поколений: К 70-летию Аллы Солломоновны Штерн. Пермь, 2011. с.121-312

[6] *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. Горизонты прикладной лингвистики и лингвистических технологий. Доклады международной конференции. – Симферополь, Изд-во: "ДиАйПи", 2007. - С. 223-224.

[7] *Allan J., Papka, R., Lavrenko V.* On-line new event detection and tracking // In SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR conference on Research and development in information retrieval. – 1998.

[8] *Падучева Е.В.* Семантические исследования. Семантика времени и вида в русском языке. Семантика нарратива. – М.: Языки русской культуры, 1996. Изд. 2-е, 2010

[9] *Ягунова Е.В., Пивоварова Л.М.* Экспериментально-вычислительные исследования художественной прозы Н.В. Гоголя // XLII Виноградовские чтения в МГУ «В.В. Виноградов о художественном тексте»: Материалы – М., 2012 (в печати)
http://webground.su/data/lit/pivovarova_yagunova/Experimentalno-vychislitelnyie_issledovaniya_prozy.pdf

[10] *Мурзин Л.Н., Штерн А.С.* Текст и его восприятие. – Свердловск : Изд-во Урал. ун-та, 1991. – 172 с.

[11] *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.

[12] *Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В*

Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие. — М.: МИЭМ, 2011. — 272 с.

Dynamic frequency features as the basis for the structural description of diverse linguistic objects

Elena Yagunova, Dmitry Lande

The paper presents an approach to studying dynamic frequency features of words for diverse linguistic objects; the purpose of the approach is to describe heterogeneous dynamic objects covering the wide range from individual texts to the news texts flow. Four groups of words are used, extracted on the basis of their dynamic frequency response (both global and local), each of which has a clearly distinct physical and linguistic nature. They correspond to diverse linguistic characteristics of objects in terms of the object structure and language features.