ИССЛЕДОВАНИЕ СТРУКТУРЫ НОВОСТНОГО ТЕКСТА КАК ПОСЛЕДОВАТЕЛЬНОСТИ СВЯЗНЫХ СЕГМЕНТОВ

Е. В. Ягунова (iagounova.elena@gmail.com)

Л. М. Пивоварова (lidia.pivovarova@gmail.com)

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

Мы рассматриваем сочетания двух и более лексических единиц, которые выделяются нами из текста на основании статистических критериев и экспериментов с информантами с учетом контекстов различного типа. Полученные списки связанных сегментов эксплицируют разные информационные структуры одного и того же текста.

Ключевые слова: списки, информационные структуры, списки сегментов, связные сегменты.

A STUDY OF THE NEWS TEXT STRUCTURE AS A CONSEQUENCE OF CONNECTED SEGMENTS

E. V. Iagunova (iagounova.elena@gmail.com)

L. M. Pivovarova (lidia.pivovarova@gmail.com)

Saint-Petersburg State University, Saint-Petersburg,
Russian Federation

The main object of this study is connected segments (collocations, compound nominations, predicative constructions, multiword expressions, etc.) extracted from the text by different statistical measures and during experiments with native speakers. This paper deals with news texts: i) 2010 news from lenta.ru (40000 texts, 9.5 million tokens); ii) a small highly homogeneous corpus that deals with some particular event: Schwarzenegger in Moscow (360 texts,110 thousand tokens) and The appointment of Sobyanin (660 texts,170 thousand tokens); iii) three individual texts about Schwarzenegger and two individual texts about Sobyanin. These texts are part

of both the small homogeneous corpus and the large news corpus. In this paper we use an open-source “Cosegment” system (http://donelaitis.vdu.lt/~vidas/tools. htm). The program cuts the text into strongly connected segments depending on the corpus. We study different types of context using overlapping corpora as the input of the system. We also compare result based on the whole corpus and on individual texts from this corpus. During the experiments with native speakers we ask 18 students to put a number from 0 to 5 between every two words in the text. 5 means that these two words are strongly connected, 0 that there is no connection at all. Then we use a cutoff 3.7 to divide a text into connected segments. Our results are the following: i) Longer connected segments are found in the more homogeneous corpus; ii) Frequent connected segments in highly homogeneous corpora (as opposed to lenta. ru corpus) are mostly predicative constructions; iii) The computer processing data are very close to the native speakers’ data; iv) Native speakers tend to extract longer segments; they also prefer predicative constructions to collocations.

Key words: words connection, connected segments, information structures, context.

1. Введение

Эта работа посвящена экспериментальному исследованию структуры текста в духе принципиальной неединственности структурированности текста. Невозможность построения единственной структуры текста, отвечающей всем возможным видам анализа, уже рассматривалась ранее (напр., Падучева 2001: 109–112; Ягунова 2008).

На предыдущих этапах исследования (см. Ягунова 2008) выборка анализируемых текстов была ограничена возможностями экспериментов с информантами, т. е. объектом исследования становились отдельные тексты. Сейчас мы пытаемся реализовать следующий виток, когда объектом исследования становятся большая текстовая коллекция объемом в миллионы словоупотреблений и тематически однородные кластеры (подколекции). В результате различных вычислительных экспериментов на основе таких коллекций мы получаем данные, с одной стороны, позволяющие соотнести особенности структуры двух разных объектов (коллекции vs. единичного текста), с другой — определить интересующие нас типы текстов (структур текстов) и, тем самым, сузить материал для экспериментальной работы с информантами. В результате мы имеем возможность наиболее тщательно исследовать роль контекста: большой коллекции текстов → тематически однородной подколлекции текстов (сюжет или кластер) → единичного текста и → минимального синтаксического контекста (подробнее см. (Ягунова 2008; Ягунова, Пивоварова 2010)). Мы в своем исследовании языка и речи идем от реализации, от имеющегося в нашем распоряжении материала.

Мы рассматриваем **все** связанные сочетания двух и более лексических единиц, которые выделяются нами из текста на основании статистических критериев и/или экспериментов с информантами. Выделяемые таким образом единицы представляют собой неоднородное множество с точки зрения соотнесенности

со словарем и/или грамматикой, номинативностью и/или предикативностью. Возникает задача интерпретации выделенных единиц. Если подойти к этой задаче не столько с точки зрения четкого разбиения на классы, но выстраивания гибкой шкалы, то типовые, или ядерные, коллокации являются сложными номинациями (единицами словаря и парадигматическими единицами). Типовые, или ядерные, конструкции находятся на другом конце шкалы и характеризуются высокой предикативностью и синтагматичностью (см. подробнее в (Ягунова, Пивоварова 2011)).

Как известно, в процедурах обработки текста происходит максимальная опора на контекст. Причем понятие «контекст» также рассматриваться в разных смыслах. Для нас контекст предполагает широкое понимание:

- минимальный контекст, в котором реализуются лексические и морфолого-синтаксические явления;
- текстовый контекст, включающий в себя фрагменты текста вплоть до текста целиком;
- контекст, предполагающий учет текстов определенного типа (заданного функционального стиля, отобранной коллекции текстов и т. д.)

Степень связанности неоднословной единицы и закономерности ее появления в тексте, по всей видимости, описываются вероятностной моделью; оценки могут быть получены лишь на основании статистических данных. Причем статистические характеристики должны описывать данные в зависимости от перечисленных выше типов контекста, т. е. контекст должен учитываться как один из параметров модели.

Таким образом, неоднословные связанные сегменты выступают, прежде всего, как структурные составляющие текста или однородных коллекций (например, сюжетов). Анализ этих структурных составляющих позволил исследовать структуру текста и/или текстов. Единицы и контекст(-ы) анализировались во взаимодействии: главным образом, контекст (равно как и коммуникативная задача) определяют выбор единиц анализа. Тематически однородная коллекция (сюжет) изучалась методами лингвистики текста (дискурса).

Основной целью работы является оценка исследовательского потенциала предлагаемого метода в изучении теоретических аспектов лингвистики текста (дискурса).

Нами оценивались следующие данные:

- данные, полученные в ходе вычислительных экспериментов:
  - список наиболее связанных n-грамм по коллекции;
  - список наиболее связанных n-грамм по подколлекции (подколлекция является тематически более однородной, чем исходная коллекция);
  - отдельные тексты, представленные в виде последовательности связанных сочетаний («сегментов» в терминологии автора программы);
- отдельные тексты, представленные в виде последовательности связанных сочетаний, полученных в ходе эксперимента с информантами.

В ходе предварительного анализа пилотного эксперимента были сформулированы некоторые **гипотезы**, которые мы проверяем в ходе данной работы:

- с увеличением степени однородности (коллекция → однородная коллекция → текст) характерными становятся более длинные n-граммы;
- с увеличением степени однородности (коллекция → однородная коллекция → текст) увеличивается число конструкций (в соотношении конструкция vs. типовая коллокация), увеличивается число предикативных сочетаний;
- набор связанных сочетаний, подсчитанных для каждого текста отдельно в ходе вычислительного эксперимента, сходен с набором сочетаний, полученных в ходе экспериментов с информантами,
- набор связанных сочетаний, выделенный в ходе экспериментов с информантами, содержит несколько больше предикативных сочетаний, чем набор связанных сочетаний, сформированный в ходе вычислительного эксперимента.

2. Методика. Гипотезы

Данное исследование предполагает сочетание вычислительного эксперимента и эксперимента с информантами. В ходе вычислительного эксперимента меры совместной встречаемости определяется на основании видоизмененной меры Дайса (Dice) (Daudaravicius 2010a):

$$Dice'(x, y) = \log_2 \left(\frac{2 * f(x, y)}{f(x) + f(y)} \right),$$

где f(x) и f(y) — частота встречаемости слов x и y в коллекции, а f(x,y) — частота совместной встречаемости слов x и y.

Процесс вычислительного эксперимента можно коротко описать следующим алгоритмом. Сначала для всех пар слов по всей коллекции считается коэффициент Дайса. Затем для каждого конкретного текста, представляющего собой цепочку слов или, вернее, цепочку пересекающихся пар (слово *x* с предшествующим словом и слово *x* с последующим словом), осуществляется «сборка» связанных сегментов. При последовательном прохождении от слова к слову в каждом тексте уже известны соответствующие значения меры Дайса для всех пересекающихся пар. На основании значений этой статистической меры слова объединяются в связанные группы с учетом ближайшего контекста (принимается решение о том, надо ли присоединить текущее слово к предыдущему). Слово не присоединяется к предыдущему, если значение коэффициента Дайса для данной пары ниже порогового, или если оно ниже, чем среднее арифметическое того же коэффициента для левой и правой пары. Во всех остальных случаях слово присоединяется. Связанный сегмент может включать не более семи слов. В результате такого вычислительного эксперимента мы получаем набор связанных сочетаний, подсчитанных для каждого текста отдельно, а затем объединенный в некое подобие частотного словаря связанных сочетаний. Программа, реализующая этот алгоритм, доступна для скачивания с сайта ее создателя: http://donelaitis.vdu.lt/~vidas/tools.htm.

В ходе интерпретации мы исходили из того, что используемая мера выделяет связанные сегменты, характеризующиеся информационной ценностью

на материале однородной коллекции текстов (ср. Daudaravičius 2010б; Daudaravičius, Marcinkevičienė 2004). Свое предположение мы проверили через сопоставление с результатами, полученными с помощью стандартных статистических мер MI и t-score, и с ключевыми словами, выделяемыми на основании коэффициента важности tf-idf (этот коэффициент позволяет оценить степень важности слова по отношению к той или иной коллекции (подколлекции)). Выдвинутое предположение об информационной значимости связанных сегментов, выделяемых с помощью меры Дайса на материале тематически однородной коллекций текстов, подтверждается в ходе предыдущих исследований с использованием меры MI (напр., Ягунова, Пивоварова 2010; Ягунова, Пивоварова 2011). При рассмотрении указанных сегментов в рамках единичных текстов (по результатам вычислительного эксперимента и эксперимента с информантами) будем называть их значимыми структурными составляющими текста (значимыми для анализа текстов).

Материалом послужили тексты и/или коллекции:

- коллекции
  - Тексты портала Лента.ру за 2010 год — 40 000 текстов общим объемом около 9,5 млн. токенов (т. е. словоупотреблений и знаков препинания);
- два сюжета (или кластера), т. е. две небольших коллекции тематически однородных текстов, полученных с помощью ресурса «Галактика Зум»[1]:
  - приезд А. Шварцнеггера в Москву — 360 текстов, около110 тыс. токенов,
  - назначение С. Собянина — 660 текстов, 170 тыс. токенов,

  все тексты кластеров берутся из новостного потока, они близки по времени появления и посвящены одному событию;
- три текста о А. Шварцнеггере (из Лента.ру, РИАН, Газета.ру) и два текста о Собянине (Лента.ру, РИАН). Эти тексты использовались в вычислительных экспериментах (с соответствующим кластером и коллекцией данного информационного источника за 2010 год в роли двух разных контекстов) и в эксперименте с информантами.

Выбор конкретных новостных текстов и сюжетов (кластеров), т. е. подколлекций, состоящих из максимально тематически однородных текстов, определялся следующими соображениями. Материал должен был обладать сравнительно четкой и простой синтаксико-семантической структурой. Отбирались кластеры сравнительно большого объема с информационно значимым сюжетом (по субъективной оценке), имеющие четко выстроенный сюжет (основное действующее лицо (или лица), основное действие, сопровождающие действующие лица и/или организации, сопровождающие действия, время, место и т. д.).

В эксперименте с информантами — эксперименте по шкалированию — приняло участие 18 студентов СПбГУ, получающих гуманитарное образование[2]. Эксперимент с информантами представлял собой оценку связности между

---

[1] Этот материал любезно предоставлен нам Александром Антоновым и Станиславом Баглеем, Галактика-Zoom: galaktika-zoom.ru, http://www.webground.su

[2] Пользуясь случаем, хотим поблагодарить Галину Доброву за помощь в проведении эксперимента.

текстоформами (пробельными словами) в шкале от 0 до 5, где 5 — соответствовало максимальной, а 0 — минимальной степени связности. В инструкции информанту предлагалось оценить «*степень связности между словами или словом и знаком препинания в шкале от 0 до 5 баллов. «0» соответствует минимальной силе связности, а «5» — максимальной силе связности. Проставьте эти баллы (от 0 до 5) во ВСЕ позиции, между ВСЕМИ словами и/или словами и знаками препинания*». Информантам отдельно не объяснялся принцип оценки связности, они должны были действовать, опираясь на интуитивные представления о связности и, конечно, на свою текстовую базу знаний. Экспериментатор не навязывает информанту предпочтение, например, синтаксического или лексико-семантического подхода, однако полученные данные позволяют судить о том, что информанты в целом справляются с поставленной задачей. Усредненные данные по группе информантов не менее 18 человек, представили непротиворечивую оценку степени связности между словами. На основании этих данных можно выстраивать сколь угодно длинные цепочки слов в соответствии с устанавливаемым пороговым значением связности. Эмпирически мы подобрали пороговое значение, равное 3,7 баллам. Если полученное число было больше, чем 3,7, пару слов рассматривали как связную, если меньше — как не связную.

Носитель языка имеет интуитивные представления о неслучайно встречающихся сочетаниях слов: текстовые базы по текстам разных функциональных стилей, по текстам разных тематик или по текстам, посвященным определенной теме. На основании этого знания адресат воспринимает каждый конкретный текст как непротиворечащий некоторой текстовой базе адресата (в качестве ее аналога при вычислительном эксперименте выступают коллекции и подколлекции текстов разной степени однородности). Тематически однородные кластеры представляли достаточно обсуждаемые события, поэтому нельзя было предположить, что информанты не знакомы с этими темами. Эксперимент проводился примерно через месяц после описываемых событий, так что эти темы не могли быть забыты.

3. Предварительные результаты

Наибольший интерес в данном докладе представлял анализ данных, полученных на материале кластеров для словоформ[3]. При интерпретации данных по рассматриваемым сюжетам мы опирались на данные, полученные на материале двух сюжетов и пяти указанных текстов, однако для иллюстрации возможностей предлагаемого метода в статье приведены результаты только двух текстов: одного текста о А. Шварценергере и одного текста о С. Собянине из «Лента.ру» 2010 года[4].

---

[3] Анализ связности для лексем (лемм) также был нами произведен, но эта тема не является предметом данной статьи. Лемматизация текстов была произведена при помощи свободно распространяемого программного обеспечения АОТ (www.aot.ru), адаптированного под наши задачи В. В. Бочаровым.

[4] В статье мы ограничиваемся новостными текстами, однако при интерпретации данных частично учитывались также результаты, полученные на материале научных

В табл. 1 представлены данные вычислительного эксперимента и эксперимента с информантами на материале сюжета и текста о А. Шварценеггере. В таблице представлены сегменты, состоящие не менее чем из трех текстоформ (слов, разделителем между которыми служат пробелы и/или знаки препинания). Полужирным шрифтом выделены те сегменты или их фрагменты, которые присутствуют как в списке, полученном в ходе вычислительного эксперимента, так и в эксперименте с информантами. В графу «Сюжет о Шварценеггере (однородная коллекция)» попала верхушка наиболее частотных связанных сегментов, упорядоченных по частоте, остальные графы (наборы) представлены в табл. 1 полностью.

Предложенная нами методика учитывает различные виды контекстов: «тематический» (сюжет) и «стилистический» (Лента.ру) (см. табл. 1). В «стилистическом» контексте существенными оказывались характерные для СМИ конструкции и обороты (например, *в настоящее время*, *со ссылкой на*), из которых нельзя сделать выводы о конкретном содержании текстов, но можно составить общее впечатление об их стилистической направленности (см. табл. 1). В «тематическом» контексте наиболее значимыми оказывались сложные номинации (*глобальное инновационное партнерство*) и предикативные конструкции, описывающие ситуацию (*только что приземлился*) (см. табл. 1). Структурные составляющие сюжета дали более полное и объективное представление о сюжете, чем структурные составляющие единичного текста. Информанты в целом выделяли более длинные сегменты, чем программа. Информанты были нацелены на описание ситуаций, они выделяли большее число предикативных сочетаний — длинные конструкции в целом более типичны, чем длинные коллокации.

Таблица 1. Связанные сегменты, состоящие не менее чем из трех текстоформ

Вычислительный эксперимент			**Эксперимент с информантами, единичный текст о А. Шварценеггера**
Коллекция (Лента.ру 2010 г)	**Сюжет о Шварценеггере (однородная коллекция)**	**Единичный текст о А. Шварценеггера**	
тем не менее	глобальное инновационное партнерство	только что приземлился	Губернатор Калифорнии Арнольд Шварценеггер
в связи с	представителей ведущих компаний	**могу дождаться встречи**	прилетел в Москву.
в 2009 году	с губернатором калифорнии	вскоре после этого	в российскую столицу
то же время	могу дождаться встречи	**ответил** калифорнийскому губернатору	Не **могу дождаться встречи** с президентом Медведевым

текстов (тематически однородная коллекция материалов конференции «Корпусная лингвистика» и 4 текста из этой коллекции).

Вычислительный эксперимент			**Эксперимент с информантами, единичный текст о А. Шварценеггера**
Коллекция (Лента. ру 2010 г)	**Сюжет о Шварценеггере (однородная коллекция)**	**Единичный текст о А. Шварценеггера**	
в настоящее время	во главе делегации	англоязычная версия твита	российский президент Дмитрий Медведев **ответил**
со ссылкой на	создать настоящий технологический бум	ответил ему взаимностью	в своем микроблоге
возбуждено уголовное дело	сфере высоких технологий	это же время	добро пожаловать в Москву
по сравнению с	только что приземлился		Жду встречи с вами
в 2008 году	тогда вам сказал		Медведев добавил микроблог
и т. д.	которые занимаются инновационными разработками		с делегацией представителей
	их российскими партнерами		он встретится с российскими министрами
	российская венчурная компания		во время посещения Медведевым
	стать мэром москвы		российский президент завел себе
	Global Technology Symposium		
	главами американских инвестиционных компаний		
	видение дальнейшего развития		
	Silicon Valley Bank		
	пост мэра москвы		
	самых разных событий происходит		
	июне этого года		
	после непродолжительной беседы		
	и т. д.		

Число пересекающихся длинных связанных сегментов, выделяемых программой и информантами, в существенной степени зависит от типа текста. Для

более динамичных сюжетов и текстов (включающих описание последовательности событий) число пересечений меньше, для более статичных — больше[5]. Это один из параметров, позволяющих оценить структуру единичного текста и текстов сюжета в целом. Нам кажется неправильным рассматривать процент совпадений между программой и информантами как меру оценки качества работы программы, поскольку информанты ничего не знали о конечных целях исследования. Оценка работы самой программы производилась ранее ее автором. В частности, в (Daudaravičius 2010b) показано, что для задачи выделения ключевых слов использование алгоритма сегментации дает улучшение F-меры на 17–27 % в зависимости от данных.

Набор длинных связанных сегментов, выделяемых информантами, на наш взгляд, может считаться самоценным для анализа структуры текста, т. к. вполне вероятно, что они отражают расстановку структурных составляющих текста, важных для восприятия (ср. идею о том, что при восприятии адресат стремится оперировать наиболее крупными оперативными единицами, напр., Грановская 1974). Продемонстрируем это на примере текста, в котором длинные связанные сегменты интерпретировались в духе гештальтпсихологии в качестве фигуры (они выделены полужирным шрифтом), а все остальные фрагменты текста рассматриваются как фон (выделены зачеркнутым шрифтом):

Губернатор Калифорнии Арнольд Шварценеггер ~~10 октября~~ **прилетел в Москву.** / ~~После прибытия~~ **в российскую столицу** ~~он сделал в своем микроблоге на Twitter соответствующую запись (Только что приземлился в Москве. Прекрасный день.~~ **Не могу дождаться встречи с президентом Медведевым**), ~~а также разместил фотографию, сделанную по дороге из аэропорта.~~

~~Вскоре после этого~~ **российский президент Дмитрий Медведев ответил** ~~калифорнийскому губернатору~~ **в своем микроблоге**: ~~@Schwarzenegger~~, **добро пожаловать в Москву**. ~~Англоязычная версия твита Медведева также содержала слова~~ «**Жду встречи с вами** ~~и вашей делегацией в @skolkovo».~~

~~Кроме того,~~ **Медведев добавил микроблог** ~~Шварценеггера в друзья. Губернатор Калифорнии ответил ему взаимностью.~~

~~Как сообщает РИА Новости, Шварценеггер приехал в Россию~~ **с делегацией представителей** ~~венчурных фондов и инновационных компаний Кремниевой долины. Планируется, что помимо президента Медведева~~, **он встретится с российскими министрами**.

~~Президент России и губернатор Калифорнии~~ **в этом году** ~~уже встречались —~~ **это произошло в июне / во время посещения Медведевым** ~~США. В это же время~~ **российский президент завел себе** ~~микроблог~~.

Набор двухсловных связанных сегментов (полученных в эксперименте с информантами), конечно, имел информационную ценность, однако количество предикативных сочетаний в нем минимально (см. табл. 2). Объединение набора двухсловных и длинных связанных сегментов «улучшает» понимание значимости визита А. Шварцнеггера для развития высоких технологий (см.

[5] По нашим предварительным данным, для научных текстов такого рода пересечений гораздо больше, чем для новостных текстов.

табл. 2), а насколько эта составляющая важна — решать адресату, т. е. тому, кто анализирует и понимает этот текст. Возможно, причина невыделения сегментов, несущих такую информацию, в том, что большинство информантов — гуманитарии, однако структура рассматриваемых текстов как минимум позволяет прочтение, в котором «*развитие высоких технологий*» является второстепенным фактом.

На материале результатов вычислительных экспериментов картина более неоднозначная. Если для кластера в целом длинные связанные сегменты информативны, то в случае единичного текста в указанном примере длинных связанных сегментов мало, мы не можем извлечь ценную информацию (понять текст) из их набора. Рассмотрим набор связанных сегментов, состоящих из 2 текстоформ, полученных в ходе вычислительного эксперимента с этим текстом (см. табл. 2). Среди них много информационно значимых единиц для описания сюжета (наименования персон, организаций, места и времени), более того — среди них неожиданно много предикативных единиц (основные выделены курсивом в табл.2). Среди них встречались цепочки (здесь и далее знак «/» показывает границу между связанными сегментами), например, ***Губернатор Калифорнии / Арнольд Шварценеггер,*** *~~Как сообщает / РИА Новости,~~* ***~~Шварценеггер приехал / в Россию~~***. При сопоставлении материалов экспериментов с информантами и вычислительного эксперимента фигура и фон — связанные сегменты, выступающие в качестве фигуры и/или фона, — могли меняться или оставаться прежними (здесь и далее полужирный и зачеркнутый шрифт соответствует выше описанному анализу результатов эксперимента с информантами). Кроме того, часто встречались случаи опущения однозначно восстановимого предлога (в примере такие предлоги заключены в скобки), напр., *~~После прибытия~~* / ***(в) российскую столицу*** / *~~он сделал / (в) своем микроблоге,~~* ***он встретится / (с) российскими министрами***. Объединение двухсловных и длинных связанных сегментов — по результатам вычислительного эксперимента — дало достаточно полное представление о структуре текста, необходимой для извлечения смысла.

Почему, если рассматривать каждый из текстов из кластера про Шварцнеггера, то длинных связанных сегментов, полученных в результате вычислительного эксперимента, практически никогда не оказывается достаточно для анализа информационной структуры этого текста? Почему для этого материала столь велико различие между набором длинных связанных сегментов, полученных в результате эксперимента с информантами и вычислительного эксперимента?

Одна из основных причин лежит в особенностях структуры анализируемого в примере текста. Телетайпный, отрывочный стиль написания большинства текстов кластера про А. Шварцнеггера (возможно, обыгрывающий общение в твиттере) характеризуется короткими структурами и навязывает короткие связанные сегменты. Характеристику анализируемого текста можно дополнить отсутствием четко выраженной композиционной структуры сюжета. Выбор примера — и сюжета, и текста как его наиболее яркого представителя — обусловил резкое различие между результатами эксперимента с информантами и вычислительного эксперимента.

Таблица 2. Связанные сегменты, состоящие из 2 текстоформ, полученные в ходе вычислительного эксперимента

губернатор Калифорнии	из аэропорта	венчурных фондов
Арнольд Шварценеггер	российский президент	инновационных компаний
10 октября	Дмитрий Медведев	
в Москву	своем микроблоге	Кремниевой долины
после прибытия	*добро пожаловать*	, что
российскую столицу	в Москву	президента Медведева
он сделал	*содержала слова*	*он встретится*
своем микроблоге	*жду встречи*	российскими министрами
соответствующую запись	вашей делегацией	
	кроме того	президент России
в Москве	добавил микроблог	губернатор Калифорнии
прекрасный день	в друзья	
президентом Медведевым	губернатор Калифорнии	этом году
	как сообщает	*уже встречались*
а также	РИА Новости	во время
разместил фотографию	*Шварценеггер приехал*	российский президент
сделанную по	в Россию	*завел себе*

Таблица 3. Связанные сегменты из текста про С. Собянина, состоящие не менее, чем из 3 текстоформ[6]

Кластер про С. Собянина (однородная коллекция)	**Вычислительный эксперимент**	**Эксперимент с информантами**
на пост мэра	**Московской городской думы**	Сергей Собянин утвержден
Московской городской думы	**проголосовали 32 депутата**	на посту мэра Москвы
проголосовали 32 депутата	участвовали 34 человека	**Московской городской думы**
тот же день	присяга нового мэра	**проголосовали 32 депутата**
губернатор Нижегородской области	тот же день	против высказались двое
нового мэра Москвы	**Как сообщалось ранее**	голосование в Мосгордуме
из 35 депутатов	**18 : 00**	**Как сообщалось ранее**
инаугурация нового мэра	избрании нового градоначальника	торжественное мероприятие планируется провести

[6] Полужирным шрифтом выделены те сегменты или их фрагменты, которые присутствуют в списках, полученных как в ходе вычислительного эксперимента, так и эксперимента с информантами.

Кластер про С. Собянина (однородная коллекция)	**Вычислительный эксперимент**	**Эксперимент с информантами**
центральном Федеральном округе	руководивший исполнительной властью	в **18:00**
кандидатуру Сергея Собянина	9 октября партия	21 октября 2010 года
на посту мэра	представила президенту четыре кандидатуры	нового градоначальника Москвы
добросовестно исполнять возложенные	список единоросов попали	исполнительной властью столицы
благополучию его жителей	**губернатор Нижегородской области**	с утратой доверия президента
участвовали 34 человека	прошлом — вице-мэр	Соответствующий указ Дмитрия Медведева
губернатором Тюменской области	**исполняющая обязанности вице-мэра**	на пост мэра Москвы
остановил свой выбор	остановил свой выбор	**губернатор Нижегородской области**
по его словам	после этого фракция	**исполняющая обязанности вице-мэра** Москвы
присяга нового мэра	из 35 мест	президент Медведев объявил
Московская городская дума	органах власти начался	аппарата правительства РФ
руководивший исполнительной властью	**городе Когалым Ханты-мансийский округа**	пообещала поддержать выбор Дмитрия Медведева
9 октября партия	**ответственные государственные посты**	в **городе Когалым Ханты-Мансийский округа**
избрании нового градоначальника	губернатором Тюменской области	в разные годы
до 2008 года	до 2008 года	занимал **ответственные государственные посты**
из 35 мест		
органах власти начался		
ответственные государственные посты		

В качестве контрпримера приведем кластер текстов о С. Собянине. В таблице 3 в первом столбце представлена верхушка наиболее частотных связанных сегментов из кластера, упорядоченных по частоте; во втором и третьем столбце результаты вычислительного эксперимента и эксперимента с информантами

на материале конкретного текста из этого кластера[7], также принадлежащего источнику Лента.ру. Наблюдается значительное сходство между наборами длинных связанных сегментов, полученных в результате эксперимента с информантами и вычислительного эксперимента. Длинные связанные сегменты, полученные в результате эксперимента с информантами, рассмотрим в силу нашего допущения как достаточные для анализа (понимания) текста.

Длинные связанные сегменты, полученные в результате вычислительного эксперимента, обладают, главным образом, одним «недостатком»: в их состав не попадают наименования персон, действующих лиц этого сюжета. Если бы мы добавили к этому набору набор двухсловных связанных сегментов или наименования персон (с элементами Ф.И.О.), то вся информация, необходимая для восстановления текста, присутствовала бы в объединенном наборе. Для рассматриваемого текста набор двухсловных связанных сегментов с элементами ФИО следующий: *Собянин утвержден, Сергей Собянин, за Собянина, Юрий Лужков, Дмитрия Медведева, помимо Собянина, Игорь Левитин, соратник Лужкова, Валерий Шанцев, Людмила Швецова, Медведев объявил, Сергее Собянине, Дмитрия Медведева, избрать Собянина, Сергей Собянин, Владимира Путина, Дмитрия Медведева, Владимира Путина*.

Полученные в наших экспериментах данные можно и нужно интерпретировать в духе исследования принципиальной неединственности структурированности текста (напр., Ягунова 2008) — в данном случае, прежде всего, информационной структурированности. Результаты вычислительного эксперимента и эксперимента с информантами эксплицируют разные информационные структуры одного и того же текста: разные варианты извлечения информации в соответствии с намерениями и возможностями адресата. Адресат (носитель языка или автомат) выделяет важные вехи в тексте на основании коммуникативной ситуации, собственных целей и задач. Разные возможности и задачи соответствуют разным коллекциям (в соответствии тематической областью коллекции и/или разной степенью однородности) или разным базам знаний информантов (степени компетентности информантов).

4. Заключение

Полученные данные не противоречат выдвинутым гипотезам. Рассматриваемая методика предоставляет исследователю возможность анализировать информационную структуру текстов, варьируя, как минимум, варианты коллекций и подколлекций и, исследуя, таким образом, тексты разных функциональных стилей (новостные, научные, официально-деловые), разных жанров, разной тематики.

В заключение хотим перечислить те сопоставительные результаты, которые не вошли в формат статьи, но, надеемся, смогут быть предметом обсуждения на конференции «Диалог 2011». Речь идет о двух видах сопоставления:

---

[7] Объем текста — 273 слова.

- с результатами, полученными на этом же материале с помощью статистических мер MI и t-score[8], а также с ключевыми словами, выделяемыми на основании коэффициента TF-IDF[9],
- с результатами, полученными по полностью аналогичной методике, на основе 4 текстов из материалов конференции «Корпусная лингвистика» (в контексте коллекции материалов «Корпусная лингвистика» за 2004–2008 годы и коллекции трудов конференции Диалог за 2003–2009 годы).

Настоящий этап исследования был посвящен экспериментальному исследованию теоретических аспектов лингвистики текста (дискурса). Надеемся, что в результате следующего этапа нам удастся получить те данные, которые будут полезны для построения модели понимания текста адресатом и для решения технологических вопросов анализа текстов и текстовых коллекций.

References

1. *Comptunig* Resourse, available at: http://donelaitis.vdu.lt/~vidas/tools.htm
2. *Daudaravicius V.* 2010. Automatic Identification of Lexical Units. Computational Linguistics and Intelligent text processing CICling-2009.
3. *Daudaravičius V.* 2010. The Influence of Collocation Segmentation and Top 10 Items to Keyword Assignment Performance. Proceedings of Computational Linguistics and Intelligent text processing CICling-2010 : 648–660.
4. *Daudaravičius V., Marcinkevičienė R.* 2004. Gravity Counts for the Boundaries of Collocations. International Journal of Corpus Linguistics, 9 (2).
5. *Granovskaia R. M.* 1974. Memory Model Perception [Vospriiatie Modeli Pamiati].
6. *Iagunova E. V.* 2008. Variability of the Strategies of Sounding Text Perception [Variativnost' Strategii Vospriiatiia Zvuchashchego Teksta].

---

[8] Меры MI и t-score содержательно различны. MI наилучшим образом позволяет определять наименования объектов (персон, организаций, географические наименования), термины и другие сложные номинации. Мера t-score, напротив, обычно позволяет выделять частотные составные слова (служебные и дискурсивные слова) и частотные конструкции (напр., *по словам*, *со ссылкой*, *сообщает РИА*) (ср. Ягунова, Пивоварова 2010). Однако для тематически однородных коллекций мера t-score выделяет информативно значимые сложные номинации, которые характеризуют коллекцию в целом и присутствуют (почти) во всех текстах коллекции (Пивоварова, Ягунова 2010; Ягунова, Пивоварова 2011).

[9] Для примера возможности сопоставления информационной важности слов текста (на примере текста про С. Собянина по отношению к коллекции лента.ру за 2010г.) приведем топ слов с наибольшими значениями TF-IDF: *Собянин, мэр, Москва, заседание, Медведев, вице-мэр, мосгордума, пост, избрание, градоначальник, 2001–2005, октябрь, Когалым, Лужков, кандидатура, Дмитрий, вице-премьер, 32, утвердить, Сергей, голосование, губернатор, аппарат, столичный, единый, четверг, Швецов, президент, Ханты-мансийский, Шанцев, Путин, новое, Левитин, тюменский, присяга, депутат, минтранс, тайный, утрата, выбор, 21, занимать, единорос.*

7. *Iagunova E. V., Pivovarova L. M.* 2010. The Nature of Collocations in Russian Language. Experimetn of Automatic Extraction and Classification on the Material of New Texts [Priroda Kollokatsii v Russkom Iazyke. Opyt Avtomaticheskogo Izvlecheiia I Klassifikatsii na Materiale Novostnykh Tekstov]. NTI, 2 (6).
8. *Iagunova E. V., Pivovarova L. M.* 2011. From Collocations to Constructions [Ot Kollokatsii k Konstruktsiiam]. Russkii Iazyk: Konstruktsionnye I Leksiko-Semanticheskie Podkhody.
9. *Manning C., Schutze H.* 2002. Collocations. Foundations of Statictical Natural Language Processing :151–189
10. *Paducheva E. V.* 2001. Statement and its Correlation with the Reality [Vyskazyvanie I ego Sootnesennost' s Real'nost'iu].
11. *Pivovarova L. M., Iagunova E. V.* 2010. Extraction and Classification of Terminological Collocations on the Material of Linguistic Scientific Texts (Preliminary Observations) [Izvlechenie I Klassifikatsiia Terminologicheskikh Kollokatsii na Materiale Lingvisticheskikh Nauchnykh Tekstov (Predvaritel'nye Nabliudeniia)]. Materialy Simpoziuma "Terminologiia I Znanie" (Proc. of Symposium "Terminology and Knowledge").
12. *Stubbs M.* 1995. Collocations and Semantic Profiles: On the Case of the Trouble with Quantitative Studies. Functions of Language, 2 (11) : 23–55.