Крылова И.В, Пивоварова Л.М., Савина А.В., Ягунова Е.В. Исследование новостных сегментов российской «снежной революции»: вычислительный эксперимент и интуиция лингвистов // Понимание в коммуникации: Человек в информационном пространстве: сб. научных трудов. В 3 тт. – Ярославль – Москва: Изд-во ЯГПУ 2012. Т.1. С. 377-382

Крылова И.В, Пивоварова Л.М., Савина А.В., Ягунова Е.В.
(СПбГУ, Университет Хельсинки)
krylova93@gmail.com, lidia.pivovarova@gmail.com,
anja.savina@gmail.com, iagounova.elena@gmail.com

ИССЛЕДОВАНИЕ НОВОСТНЫХ СЕГМЕНТОВ РОССИЙСКОЙ «СНЕЖНОЙ РЕВОЛЮЦИИ»: ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ И ИНТУИЦИЯ ЛИНГВИСТОВ

Irina Krylova, Lidia Pivovarova, Anna Savina & Elena Yagunova (St.-Petersburg State University, University of Helsinki)

COMPARATIVE INVESTIGATION OF NEWS SEGMENTS «RUSSIAN SNOW REVOLUTION»: A COMPUTATIONAL EXPERIMENT AND INTUITION OF LINGUISTS

News segment, «Russian snow revolution», experimental method, computational linguistics, intuition of linguists, keywords, media resources, interpretation, degree of theme importance

В статье предлагается методика исследования новостных документов временного сегмента Российской «Снежной революцией», в которой сопоставляются данные вычислительных экспериментов и интуиции лингвистов («Словарь XXI века» (рук. А.В.Михеев)).

We discuss a research methodology for news texts – the time segment of the Russian "Snow Revolution" – with comparing the results of calculations and linguists' intuition ("Dictionary of XXI century" (supervised by Aleksey Miheev)).

Мы живем в то время, когда лингвист, работающий с новостными текстами, не может не заинтересоваться спецификой разных современных сегментов (срезов) по данным СМИ. Для последних временных сегментов наблюдается очень активный всплеск новых тем и, соответственно, слов, маркирующих эти темы; наблюдается словотворчество, зачастую новые слова маркируют не только несколько более свободные блоги, но более традиционные источники СМИ. Конечно, для полноценного анализа

специфики тех или иных фрагментов желательно учитывать не только СМИ, однако даже анализ СМИ дает интереснейшие результаты.

Для нас представляют максимальный интерес:

- учет особенностей стиля и иерархии тем в разных СМИ,
- насыщенность сегментов разных СМИ специфическими тематическими маркерами,
- соотнесение вычислительного портрета рассматриваемых сегментов (их объективированных портретов) с тем, что активно обсуждается лингвистами (прежде всего проектом «Словарь XXI века» (рук. А.В.Михеев)).

В качестве анализируемого сегмента нами взят сегмент, близкий по времени к тому, что принято называть Российской «Снежной революцией». Конечно, в этом временном сегменте представлены разные темы, не только темы, относящиеся к тематике «Снежной революции». Однако, очевидно, что важность названной тематики не может не найти отражение в текстах, что может и должно быть выявлено как на основании интуиции лингвиста, так и на основании вычислительных экспериментов.

Материал

Исследовательские корпуса текстов пяти российских источников СМИ (периода от выборов в Думу до выборов президента): газеты «Независимая газета» (1 декабря 2011 г. – 7 марта 2012 г.), «Российская газета» (1 декабря 2011 г. – 8 марта 2012 г.), порталов «Лента.ru» (1 декабря 2011 г. – 10 марта 2012 г.), «RBC.ru» (1 декабря 2011 г. – 10 марта 2012 г.) и РИА Новости (1 декабря 2011 г. – 10 марта 2012 г.). Далее эти миникорпуса в тексте называются документами.

Методика вычислительного эксперимента

Предварительный экспресс-анализ основывался на рассмотрении коллокаций (биграмм) по каждому из корпусов, полученных с помощью

статистической меры MI (или PMI)[1], прежде всего, анализировались «топы» списков из 300 биграмм с максимальными значениями данной меры (см. подробнее Ягунова, Пивоварова 2010).

Полученные списки – как результат экспресс-анализа искомого новостного сегмента разных СМИ – сравнивали по нескольким параметрам:

- наименования персон (ФИО персон),
- географические наименования,
- другие сложные номинации.

В теории информационного поиска признано ранжирование весов слов по классическому критерию Солтона TF IDF (Salton, Buckley 1988), 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. Предлагаемый в данной работе подход близок этой идеологии: мы вычисляем TF для документа (т.е. для миникорпуса конкретного источника[2]) и TF для коллекции (множества текстов этого источника за достаточно большой временной промежуток). Вес слова определяется на основании соотношения TFдокумента и TFколлекции. До некоторой степени этот подход соотносим также с критериями локальной и глобальной частот (Ландэ и др. 2007)[3]. «Глобальная частота встречаемости – абсолютная частота встречаемости слова в анализируемом объекте. <…>

---

[1] *MI* (mutual information, коэффициент взаимной информации) сравнивает зависимые контекстно-связанные частоты с независимыми, как если бы слова появлялись в тексте совершенно случайно:

$MI=log_2 \frac{f(n,c)\times N}{f(n)\times f(c)}$, где MI – объем информации; n – n-е ключевое слово; c – коллокат; f(n,c) – абсолютная частота встречаемости ключевого слова *n* в паре с коллокатом *c*; f(n), f(c) – абсолютные частоты ключевого слова *n* и слова *c* в корпусе; N – общее число словоформ в корпусе (Stubbs M. 1995).

С точки зрения теории вероятности, мера MI является способом проверить независимость появления двух слов в тексте — если слова полностью независимы, то вероятность их совместного появления равна произведению вероятностей появления каждого из них, т.е. произведению частот (использование абсолютных частот вместо относительных увеличивает значение MI для всех коллокаций в корпусе на константу, однако не меняет ее вероятностного смысла).

[2] Все тексты временного сегмента снежной революции этого источника СМИ

[3] В (Ланде и др. 2007) исследуется и обосновывается алгоритм выявления документов «New Event Detection» (Allan et. al. 1998), который получил большую популярность только в последнее время

Локальная частота встречаемости – абсолютная частота встречаемости слова в окне наблюдения из K слов» (Ягунова, Ландэ 2012).

Целями работы является:

- выделение в ходе вычислительных экспериментов слов с высоким коэффициентом «ключевости», т.е. слов, характеризующих сегмент снежной революции, соотносимых с подачей материала в каждом из рассматриваемых источниках СМИ;
- сопоставление этих ключевых слов с теми, что выделяются лингвистами (методом экспертного голосования) в ходе работы над проектом «Словарь XXI века» (рук. А.В.Михеев).

Результаты. Обсуждение результатов

Приведем некоторые предварительные результаты, кратко остановимся на пересечениях номинаций (номинации из топа в 300 биграмм) по различным источникам.

Наименования персон (доля ФИО персон от всех биграмм топа): 50% – «Лента.ru», 39% «Российской газете», 36% – «Независимая газета», 27% – «RBC.ru». Пересечений очень мало, зачастую попадание того или иного наименования персоны требует особого обсуждения. Примерами пересечений служат «*Башар Асад*», «*Рашид Нургалиев*», «*Ильмах Алиев*» и т.д., т.е. эти пересечения не соотносятся с какой-либо единой темой.

Аналогично обстоит дело с географическими наименованиями (приводятся от максимального значения меры и далее по убыванию): напр., «*Корейский полуостров*», «***Персидский залив***», «***Саудовская Аравия***» – «Независимая газета»; «***Персидский залив***», «*Мальвинские острова*» – «RBC.ru»; «*Набережные Челны*», «***Саудовская Аравия***», «*Курильская гряда*» – «Lenta.ru»; «*Вышний Волочек*», «***Саудовская Аравия***», «*Нижний Тагил*» – «Российская газета».

Остальные биграммы (топ в 100 биграмм) 4 информантами в ходе миниэксперимента были разделены на две группы: правильные биграммы (они безоговорочно являются сложными номинациями или языковыми

клише) и ожидаемые биграммы (обладающие меньшей – но существенной для носителя языка – целостностью).

Топ-список «правильных биграмм» (отнесенные к тематике «снежной революции» выделены п/ж шрифтом или подчеркиванием (более широкая тематика)): Независимая газета – *волшебный пендель, зубная щетка, спиртные напитки, чрезвычайная ситуация, тройская унция, ассоциированное членство, дальнее зарубежье, камень преткновения, кассационная жалоба,* ***«левый фронт»****, скорая помощь, глобальное потепление, круглый стол, мультимедиа арт, птичий грипп,* ***домашний арест****,* ***арабская весна****,* *денежное довольствие*, *правоохранительные органы*, *смертная казнь,* *«эффективная политика»*;
РосБизнесКонсалтинг – *10-дневное безделье, ввозная пошлина, воинские почести, колорадский жук, Бельгийский лежье, желудочно-кишечный тракт, инфекционные болезни, однополые браки, покорный слуга, почетная грамота, рейдерские захваты, стихийные бедствия, фигурное катание, черепно-мозговая травма, вечная мерзлота, горловое пение,* ***оранжевая революция****,* *святейший владыка*;
Lenta.ru – *венские балы, дальнее зарубежье, ложный донос, магнитное поле, нацистское приветствие, стихийные бедствия, божьи коровки, Бурановские бабушки, квант милосердия, кишечная палочка,* ***открепительные удостоверения****, параноидальная шизофрения, потерянные рай, процентное соотношение, Сосновый Бор, тасманийский дьявол, физическое насилие, эпилептический припадок, гаагский трибунал, гречневая крупа, латексные перчатки*;
Российская газета – *козел отпущения, платежное поручительство, плечевой сустав, сахарная свекла, адронный коллайдер, бородинское сражение, букмекерские конторы, денежное довольствие, желудочно-кишечный тракт, завидной регулярностью, Чартова дюжина, обвинительная связка, терапевтическая эквивалентность, младых ногтей, рейдерские захваты, смертная казнь, муравьиная кислота*.

С помощью сопоставительных частотных критериев, основанных на мере TF, были выделены слова, для которых TFколлекции относительно небольшая, а TFдокумента – высокая. Именно эти слова в первую очередь полагаем соотносимыми с тематикой «снежной революции». Их дополняет второй класс слов: слова, для которых и TFколлекции, и TFдокумента относительно небольшие[4] (ср. Ягунова, Ланде 2012). Таким образом рассматривается два уровня ключевости.

[4] Первый класс ключевости : TFдокумента не менее 25, TFколлекции не более 1000; второй класс ключевости: TFдокумента менее 25, TFколлекции не более 1000

Приведем три топа для первого класса ключевости, в порядке убывания TFдокумента): Независимая газета – *избиратель, предвыборный, голосование, штаб, сторонник, парламентский, минувший, участок, ЦИК, форум, Сирия, евро, накануне, полиция, опубликовать, палата, налог, Евгений, голосовать, фон, Родина, понедельник, Киргизия, сеть, суббота, Кремль, вырасти, Приднестровье, оппозиционный, Интернет, честной, зарплата, МИД, Яблоко, назначить, Евросоюз, протестный, Молдавия, законопроект, средний, верховный, организатор, Болотная, Горбачев, премия, Интерфакс, вторник, доклад, Казахстан, законодательство, Прохоров, ставка, проголосовать, активность, санкция, СНГ, Зюганов, модернизация, фестиваль*;
Лента – *Сирия, ЦИК, Джиоев, избиратель, честной, Коммерсантъ, подпись, Ким, справедливый, Болотная, КПРФ, релиз, наблюдатель* (см. табл. 1), *санкция, сирийский, фальсификация, Навальный, шествие, проспект, редактор, Прохоров, Осетия, регистрация, Жанаозен, Чен, связать, ЛДПР, штаб, Яблоко, революция, фон, увольнение, согласовать, слух, давление, выдвинуть, атака, резолюция, Илья, блогер, Асад, регулярный, работник, Центризбирком, хакер, телевидение, арабский, предвыборный, Евросоюз, проголосовать, Тегеран-32, ресторан, шайба, запрос, нижний, приостановить, писатель, Египет, спикер, кандидатура*;
РИАН – *Джиоев, Болотная, Центризбирком, Прохоров, патриот, жеребьевка, ГУМВД, Избирком, эсер, депутатский, думский, сирийский, понижение, митинговать, Прохор, пикет, Косовский, сизый, ледокол, фальсификация, Приднестровье, Навальный, оппозиционер, рейтинговый, ЕЦБ, честной, Жириновский, недовольный, шествие, санкционировать, баллотироваться, АТЭС, Ливия, Рождественский, пересмотр, Зюганов, ГД, баскетболист, Миронов, текстовой, Цхинвали, протестный, Батурин, понизить, атлант, евролига, выдвижение, Смирнов, колонна, регулятор, Барс, союзный, лозунг, Мишарин, удостоверение*.

Сопоставление топов (ключевых слов, а также правильных биграмм) позволяет оценить иерархию тем рассматриваемых сегментов, достаточно существенную роль тематики «Снежной революции» в заданном временном сегменте и тот новостной контекст, в который эта тематика была помещена.

Таблица 1 приведена на основе ресурса http://www.facebook.com/groups/slovargoda/doc/449684175045899/ (доступ 20/08/2012).Она дополнена поиском соответствий по анализируемым сегментам 3 источников СМИ (без указания времени появления). Серый фон выделяет слова, зафиксированные в словарях применительно к более позднему периоду, п/ж шрифт отмечает наличие соответствия, без подчеркивания – слова, для которых TFколлекции относительно небольшая,

а TFдокумента – высокая, с подчеркиванием – слова, для которых и TFколлекции, и TFдокумента относительно небольшие.

Таблица 1. Универбы-2012: первая двадцатка по числу лайков (январь – май)

Слово	(месяц)	Есть ли соответствие в корпусах		
		Независимая газета	Лента	РИАН
Кощунницы	(апрель)			
Интернет-демократия	(май)	Интернет		
Карусель	**(март)**	**Карусель**	**Карусель**	**Карусель**
Печеньки	(март)			
Пуськи	(март)			
Наблюдатель	**(март)**	**Наблюдатель**	**Наблюдатель**	наблюдательный
ОккупайАбай	(май)			
Болотироваться	(январь)			
Листалка	(февраль)			
Потупчики	(февраль)			
Автозак	**(май)**	**Автозак**	**Автозак**	**Автозак**
Соскайпиться	(февраль)			
Термобелье	(февраль)			
Фонтан	**(март)**	**Фонтан**	**Фонтан**	**Фонтан**
Чуровщина	(март)	Чуров	Чуров	Чуров
Коворкинг	(январь)			
Пропаганки	(январь)			
Путинг	(январь)			
Рукопожометр	(февраль)			
Стояние	**(апрель)**	**Стояние**	**Стояние**	**Стояние**
Фитюлька	(апрель)			

В сегменте «Независимой газеты» представлено несколько больше слов из списка «Универбы-2012: первая двадцатка по числу лайков (январь – май)», прежде всего, это проявляется в том, что для «Независимой газеты» эти слова принадлежат к наиболее яркому классу слов, для которых TFколлекции относительно небольшая, а TFдокумента – высокая (без подчеркивания).

Анализируемые в вычислительном эксперименте документы, относятся к временному сегменту до 7-10 марта, однако в них выделяются и те ключевые слова, которые оцениваются лингвистами как маркеры апреля и марта.

Часть слов списка не попало в интересующие нас классы ключевости (двух уровней), т.к. принадлежат стилю блогов, (еще?) не проникшему в документы этих СМИ (*Бо**о**лотироваться*, *Пропаганки*, *Путинг* и т.д.).

В будущем мы планируем соотнести «реальное» и «ощущаемое» время появления в рассматриваемых СМИ тех или иных ключевых слов (т.е. анализ будет учитывать периодизацию и «локальную частоту встречаемости» этого слова в рамках рассматриваемого периода). Ну и, конечно, рассматриваемый временной сегмент будет расширен до периода инаугурации президента.

Литература

1. *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.
2. *Ягунова Е.В., Ландэ Д.В.* Динамические частотные характеристики как основа для структурного описания разнородных лингвистических объектов // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.
3. *Ягунова Е.В., Пивоварова Л.М.* Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов – Сб. НТИ, Сер.2, №5. М., 2010
4. *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.
5. *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval. Information Processing and Management, 1988. – № 24(5). – P. 513-523.
6. *Stubbs M. 1995.* Collocations and semantic profiles: on the case of the trouble with quantitative studies. // Functions of language 2:11, 23-55, Benjamins, 1995.

7. Эл. ресурс Михеев А.В. Словарь года. Универбы-2012: первая двадцатка по числу лайков (итоги первого полугодия) http://www.facebook.com/groups/slovargoda/doc/449684175045899/ (доступ 20/08/2012)