Корпусные исследования

Корпусные методы исследований языковых процессов

Даша Попова

Что такое корпус?

- ***corpus*** (Latin) – тело, плоть, структура, объединение, (позднее) коллекция текстов одного автора или одной направленности;

- ***лингвистическим корпусом*** называют совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой;

  иногда корпусом, или ***«корпусом первого порядка»***, называют просто любое собрание текстов, объединённых каким-то общим признаком (языком, жанром, автором, периодом создания текстов).

https://ru.wikipedia.org/wiki/Корпусная_лингвистика

Характеристики корпуса: репрезентативность

- Тексты, входящие в корпус, должны быть собраны по определенным принципам, чтобы представлять определенный пласт языка или весь язык в определенный период времени. Это параметр называется **репрезентативность** (англ. *representativeness*).

- **Репрезентативность** – свойство корпуса, заключающееся в статистически достоверном представлении языка или его части и достигаемое за счет необходимого объема и жанрового разнообразия текстов.

- **Представительная, или репрезентативная, выборка** (англ. *representative sampling*) – такой объем материала, увеличение которого уже почти никак не повлияет на распределение единиц.

(Копотев 2014: Глава 1)

Характеристики корпуса

- ***сбалансированность*** (англ. *balance*) -- этот параметр определяет, насколько равномерно представлены тексты разных типов (например, письменные и устные);

- ***объём корпуса***;

- ***реализация корпуса –*** сейчас дефолт: электронная форма;

- ***разметка*** (аннотация, англ. *annotation*) – это введенная автоматически или вручную лингвистическая или метатекстовая информация обо всех выбранных единицах корпуса: тексте, предложении, морфеме, звуке и т. д.

(Копотев 2014: Глава 1)

Зачем нужны корпуса?

- лингвистическое сафари;
- большие данные (при большом объёме корпуса);
- возможность многократного использования созданного корпуса для решения различных (социо) лингвистических задач.

История корпусной лингвистики

- в современном понимании дисциплина сложилась в 60-80е гг. 20в.;
- доцифровая корпусная лингвистика (примеры):
  - **грамматика Па́нини** (Pāṇini), VI-IV вв. до н.э. – грамматика санскрита, передавалась устно, 3959 стихов, основана на корпусе ведических текстов;
  - с 13в. списки слов из Библии с указанием стихов -- **симфонии**, или **конкорда́нции;**
  - создание словарей;
  - **интерес к реальному**, а не специально сконструированному, **языковому материалу**: в «Modern English Grammar on Historical Principles» (1909–1949) Отто Есперсена список источников занимает 40 страниц.

Два направления корпусной лингвистики:

- создание корпусов;
- корпусные исследования, другими словами, исследования языка при помощи корпусных методов.

Ограничения на использование корпусных методов

- нет отрицательного лингвистического материала;
- нет грамматически возможного, но прагматически не встречающегося материала;
- ошибка писца или редкость?
- сложности с аннотацией

Poverty of the **stimulus** (POS)

- Poverty of the stimulus (POS) is the argument from linguistics that children are not exposed to rich enough data within their linguistic environments to acquire every feature of their language. This is considered evidence contrary to the empiricist idea that language is learned solely through experience.

- The POS is often used as evidence for universal grammar. This is the idea that all languages conform to the same structural principles, which define the space of possible languages.

- Both poverty of the stimulus and universal grammar are terms that can be credited to Noam Chomsky. Chomsky coined the term "poverty of the stimulus" in 1980, however he had argued for the idea since his 1959 review of B.F. Skinner's *Verbal Behavior*.

https://en.wikipedia.org/wiki/Poverty_of_the_stimulus

Poverty of the **stimulus** (POS)

(1) a. The boy is happy.
b. Is the boy happy?

(2) a. The boy who is smiling is happy.
b. Is [the boy who is smiling] _ happy?
c. *Is [the boy who _ smiling] is happy?

(3) $Rule_1$: move the first auxiliary of the declarative sentence in initial position

(4) $Rule_2$: move the main auxiliary of the declarative sentence in front of the subject NP

Примеры корпусов

- Национальный корпус русского языка (НКРЯ)
  https://ruscorpora.ru/
  включает диалектные, литературные, исторические, современные, письменные, устные … тексты; лингвистическая разметка представлена морфологической, синтаксической и семантической аннотациями

- Мангеймский корпус немецкого языка
- Корпус современного американского английского языка COCA
- Британский национальный корпус BNC
- Treebank
- Switchboard
- Корпус сновидений: http://spokencorpora.ru/
- Корпус городских диалектов
- WALS: https://wals.info/
- …
- CLARIN (www.clarin.eu/) и ELRA (http://www.elra.info/) – каталоги корпусов

Классификация корпусов

- **1. Язык текстов**
- «Самое простое деление корпусов предполагает выделение **одноязычных** (англ. *monolingual*), то есть содержащих тексты на одном языке, и **многоязычных** (англ. *multilingual*). Многоязычные корпуса в свою очередь могут состоять из разных текстов, возникших, например, в ситуации многоязыкового общения, или одинаковых текстов, переведенных на разные языки. Последние представлены в виде **параллельного корпуса** (англ. *parallel corpus*), в котором тексты на разных языках связаны на уровне предложений или абзацев (**выравнивание,** англ. *alignment*). Особым типом корпуса является **сравнительный корпус** (англ. *comparable corpus*), в котором по определенным одинаковым критериям собраны тексты на разных языках или вариантах языка.
- Самая переводимая книга – Библия. Число языков, на которые она переведена целиком или частично, приближается к трем тысячам. Параллельный корпус переводов Библии уже много лет создается в Университете Мэриленд (США) и пока не закончен.»

Копотев 2014: Глава 4

Классификация корпусов

2. Тип текстов

а) письменные тексты,
б) устные (аудиозаписи и видеозаписи),
в) смешанные (мультимодальные)
…

3. Жанры текстов

а) литературные,
б) диалектные,
в) разговорные,
г) публицистические,
д) исторические,
е) корпуса второго языка (ученические и т. п.).

Копотев 2014: Глава 4

Классификация корпусов

4. **Тип данных:**
   а) полнотекстовые,
   б) фрагментированные тексты:
   1) n-граммный,
   2) конкордансный.

N-граммы (англ. *n-grams*) – цепочки, состоящие из идущих подряд двух, трех, четырех и т. д. токенов (их называют, соответственно, биграммы, триграммы, 4-граммы и т. д.).

https://books.google.com/ngrams

Конкордансом (англ. *concordance*) в корпусной лингвистике называют список найденных примеров (вхождений) нужного токена или леммы в минимальном контексте. Обычно такой контекст представляет собой фрагмент из нескольких единиц слева и справа.

Токен (token) – словоформа, лемма (lemma, type) – словарная форма

Копотев 2014: Глава 4

Классификация корпусов

5. **Типы разметки:**
   а) неразмеченные,
   б) размеченные (аннотированные), с типами разметки:
   1) метатекстовая (жанр, время создания текста и т. д.),
   2) лингвистическая:
      - фонетическая,
      - просодическая,
      - морфологическая (полная или только частеречная),
      - словообразовательная,
      - синтаксическая,
      - семантическая
      - и др.,
   3) экстралингвистическая (маркировка эмоций, жестов и т. п.).

Копотев 2014: Глава 4

Классификация корпусов

6. **Объем данных:**
   а) представительный корпус (национальный),
   б) иллюстративный

7. **Тип доступа:**
   а) свободно распространяемый,
   б) академическая лицензия,
   в) ограниченный доступ.

8. **Страна создания и авторские права**.

Копотев 2014: Глава 4

Аннотация

Аннотация – это приписанная всем единицам выбранного уровня (текст, предложение, словоформа и т. д.) соответствующая лингвистическая информация.

Например, морфологически аннотированный корпус содержит морфологический разбор частей речи.

Коптев 2014: Глава 5

Аннотация

Принципы Лича:

- «Разметка должна основываться на доступной для пользователя в виде руководства или инструкции схеме анализа, в которой введение каждого параметра должно быть мотивировано.

- Разметка общедоступного корпуса должна быть «теоретически нейтральна», то есть схема разметки по возможности должна не разрывать с традицией, а опираться на знакомую всем систему понятий. Если корпус предназначен не для конкретного проекта, то при его разметке стоит избегать пусть и строгих, но авторских, не общепринятых классификаций, которые требуют предварительного знакомства с той или иной теорией.

- Должно быть ясно, кто и как разрабатывает схему аннотации и каковы ограничения, например юридические или технические, при пользовании корпусом.»

Коптев 2014: Глава 5

Аннотация

Адам Килгарифф (Adam Kilgariff) выделил следующие этапы развития автоматического анализа текста:

- **токенизация** (англ. *tokenization*): выделение в текстовом потоке минимальных фрагментов для последующего анализа (в корпусной лингвистике их принято называть **токены** (англ. *token*);
- **лемматизация** (англ. *lemmatization*): определение для всех токенов их начальной формы (точнее **леммы** (англ. *lemma*);
- **частеречная разметка** (англ. *POS tagging*): определение части речи каждого слова;
- **полная морфологическая разметка** (англ. *full morphological tagging*): приписывание словоформе морфологических признаков;
- **синтаксическая разметка, или парсинг** (англ. *parsing*): приписывание определенных синтаксических признаков слову или сочетанию слов;
- **семантическая разметка** (англ. *semantic annotation*): включение лексемы в определенный лексико-семантический класс;
- **создание семантических сетей** (англ. *semantic network, frame network*): маркировка семантических связей между лексемами.

Коптев 2014: Глава 5

Аннотация

- ★ WordNet: http://wordnetweb.princeton.edu/perl/webwn
- ★ FrameNet: https://framenet.icsi.berkeley.edu/framenet_search
- ★ Прагматикон: https://pragmaticon.ruscorpora.ru/
- ★ Малые языки Сибири: динамика языковой ситуации: https://socio-siberian-lang.iling-ran.ru/
- ★ Корпус русской речи Карелии: http://lingconlab.ru/karelrus/#!/
- ★ Корпус дагестанского русского: http://parasolcorpus.org/dagrus/#!/

Кошмар для аннотаторов: омонимия и полисемия

- **омонимия** – случайное совпадение слов или частей слов по звучанию и написанию;

- **полисемия** – наличие у слова нескольких, возможно, исторически связанных, значений.

НКРЯ

- Сколько падежей и какие в аннотации НКРЯ?
- Приведите аргументы в пользу мужского/среднего рода «кофе».
- Обоснуйте своё решение с помощью корпусных данных.
- «творожник/сырник», «поребрик/бордюр» -- что мы можем узнать из НКРЯ и что не можем.
- «воскресение» и «воскресенье» -- диахрония.

Web as a corpus

- ✔ большие данные (big data);
- ✔ данные электронного модуса – чаты, посты;
- ✔ возможность извлекать некоторые метаданные, которых нет в созданных корпусах;
- ✔ возможность извлекать необычную аннотацию;

НО! нужно придумать, как, что и откуда извлекать;

- ✔ это уже не сафари, а джунгли

- ✔ crowdsourcing

Web as a corpus: crowdsourcing

- https://www.urbandictionary.com/

- De Marneffe et al.: 177 Mechanical Turk workers were asked to decide whether the bold-faced event described in the sentence did (or will) happen

Web as a corpus: crowdsourcing

- Standard semantic analysis (SSA) – non-factive verbs (*say*, *think*, *report*) do not convey an assumption on whether the bold-faced event described in the sentence did (or will) happen
- Prediction of the SSA:
- Uu – unknown

- De Marneffe et al.:
- Results:
- CT+ certainly true
- PR+ probably true
- PS+ possibly true

a. Magna International Inc.'s chief financial officer **resigned**, the company said. [CT+: 10]

b. In the air, U.S. Air Force fliers say they have **engaged** in "a little cat and mouse" with Iraqi warplanes. [CT+: 9, PS+: 1]

c. Merieux officials said that they are "highly confident" the offer will be **approved**. [PR+: 10]

d. U.S. commanders said 5,500 Iraqi prisoners were taken in the first hours of the ground war, though some military officials later said the total may have **climbed** above 8,000. [PS+: 7, PR+: 3]

Экспрессивные указательные местоимения

- Acton and Potts, 2013; Potts, 2013

До Трампа: Sarah Palin

“Americans are cravin’ that straight talk”

Acton and Potts, 2013; Potts, 2013

Бурная реакция на ее речь

FoxNews.com:

- *“We feel like she talks like we do.”*
- *“She talked like real people to real people”*

Huffington Post:

- *“illusion of straight-talking”*
- *“pseudo-folksiness and fundamental dishonesty”*

Acton and Potts, 2013; Potts, 2013

Дебаты 2008

- Joe Biden, Palin's opponent:

"We should be helping them build schools to compete for those hearts and minds of the people in the region"

Acton and Potts, 2013; Potts, 2013

И все, все, все:

- В полиции: *make that phone call right now*
- Стюард(есса): *get those bags under that seat in front of you*
- Инструктор йоги: *get that left arm up over that head*

Acton and Potts, 2013; Potts, 2013

Экспрессивные употребления указательных местоимений

- давно известны лингвистам

- есть и в русском:
  - Смех смехом, а в полицию уже обратилось более двадцати горожан, которые уверяют, что после встречи с этим "похитителем человеческих душ"(так окрестили маньяка журналисты) у них чего-то не хватает. Ох, и мнительные эти американцы! У нас бы на такого маньяка народ молился. Лос-анджелесская полиция злоумышленника усиленно разыскивает, но пока безуспешно. [Похититель душ // «Криминальная хроника», 2003.06.24]
  - *эти дети мне уже надоели*
  - *Уберите уже эти ноги из прохода!*

Acton and Potts, 2013; Potts, 2013

Корпусный подход

Эти утверждения импрессионистические

Можем ли мы их квантифицировать?

Acton and Potts, 2013; Potts, 2013

Experience Project: признания

Acton and Potts, 2013; Potts, 2013

Experience Project: признания

Confession:	I bought a case of beer, now I'm watching a South Park marathon while getting drunk :P
Reactions:	*Sorry, hugs*: 2; *You rock*: 3; *Teehee*: 2, *I understand*: 3; *Wow, just wow*: 0
Confession:	subconsciously, I constantly narrate my own life in my head. in third person. in a british accent. Insane? Probably
Reactions:	*Sorry, hugs*: 0; *You rock*: 7; *Teehee*: 8; *I understand*: 0; *Wow, just wow*: 1
Confession:	I really hate being shy . . . I just want to be able to talk to someone about anything and everything and be myself. . . That's all I've ever wanted.
Reactions:	*Sorry, hugs*: 1; *You rock*: 1; *Teehee*: 2; *I understand*: 10; *Wow, just wow*: 0;

Acton and Potts, 2013; Potts, 2013

Experience Project: признания

10 Reactions

Figure: EP reaction icons.

Sorry, hugs	sympathy
You rock	cheering, supportive
Teehee	amused
I understand	solitary
Wow, just wow	shock

Table: Interpreting the icons.

Acton and Potts, 2013; Potts, 2013

Experience Project: признания

Texts	140,467
Words	21,518,718
Vocab	143,712
Mean words/text	153.19

(a) The overall size of the corpus.

Category	Reactions
Sorry, hugs	91,222 (22%)
You rock	80,798 (19%)
Teehee	59,597 (14%)
I understand	125,026 (30%)
Wow, just wow	60,952 (15%)
Total	417,595

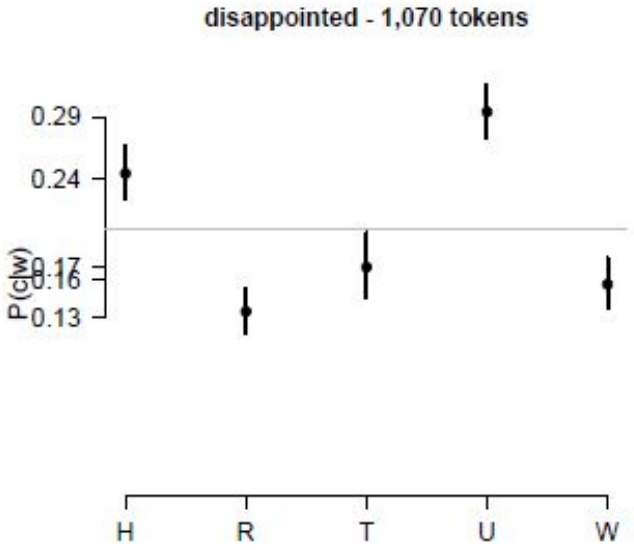
(b) All reactions.

Table: In general, reader reactions are sympathetic and supportive.

Acton and Potts, 2013; Potts, 2013

Считаем и изображаем

Cat.	Count	Total	$\Pr_{EP}(w\|r)$	$\Pr_{EP}(r\|w)$
Sorry, hugs	1167	18038374	0.00006	0.26
You rock	520	14066087	0.00004	0.15
Teehee	300	8167037	0.00004	0.15
I understand	1488	20466744	0.00007	0.29
Wow, just wow	473	12550603	0.00004	0.15



Acton and Potts, 2013; Potts, 2013

Указательные местоимения

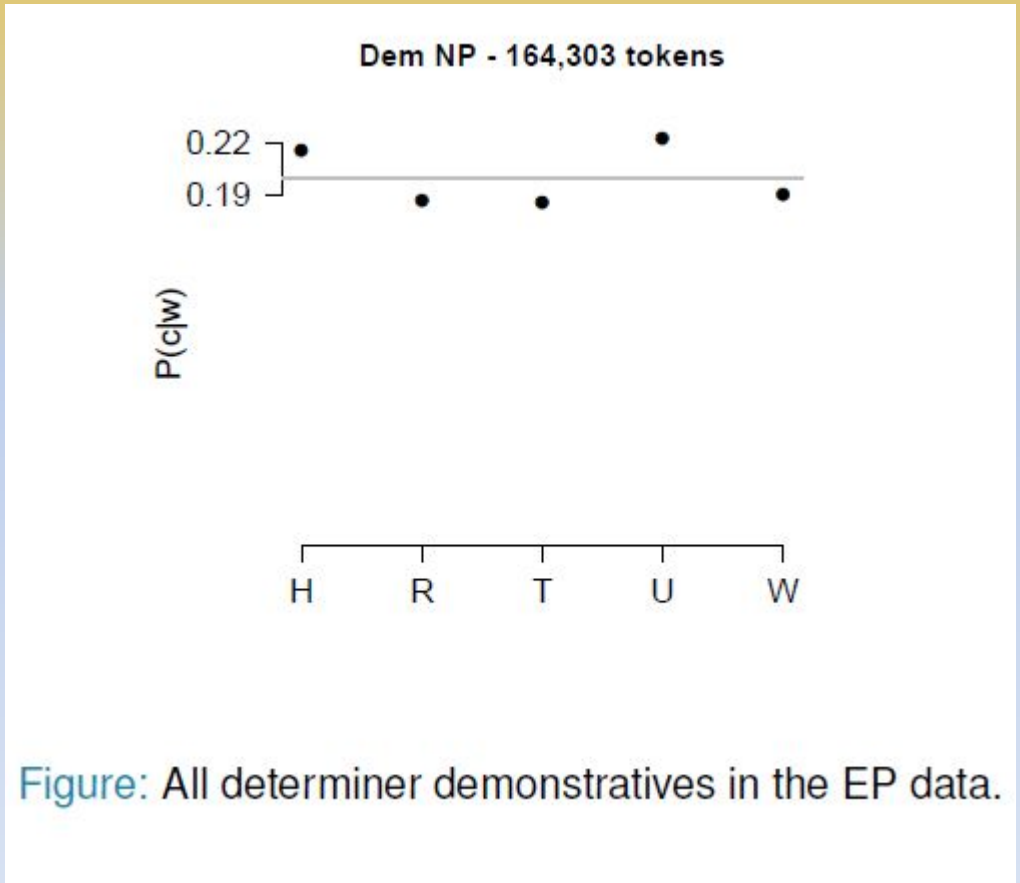


Figure: All determiner demonstratives in the EP data.

Acton and Potts, 2013; Potts, 2013

Примеры: слова, вызывающие симпатию

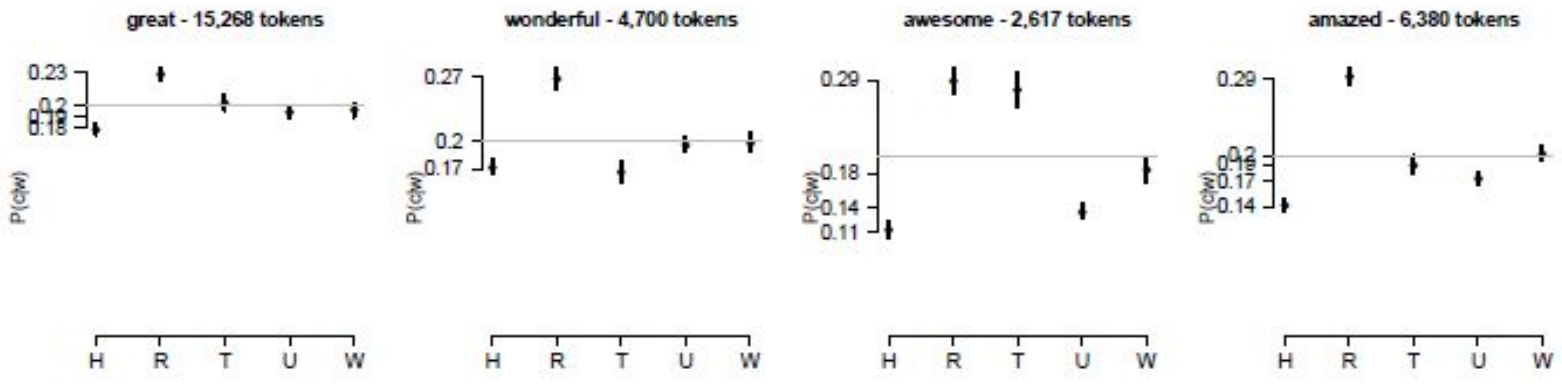


Figure: Words eliciting predominantly ‘you rock’ reactions.

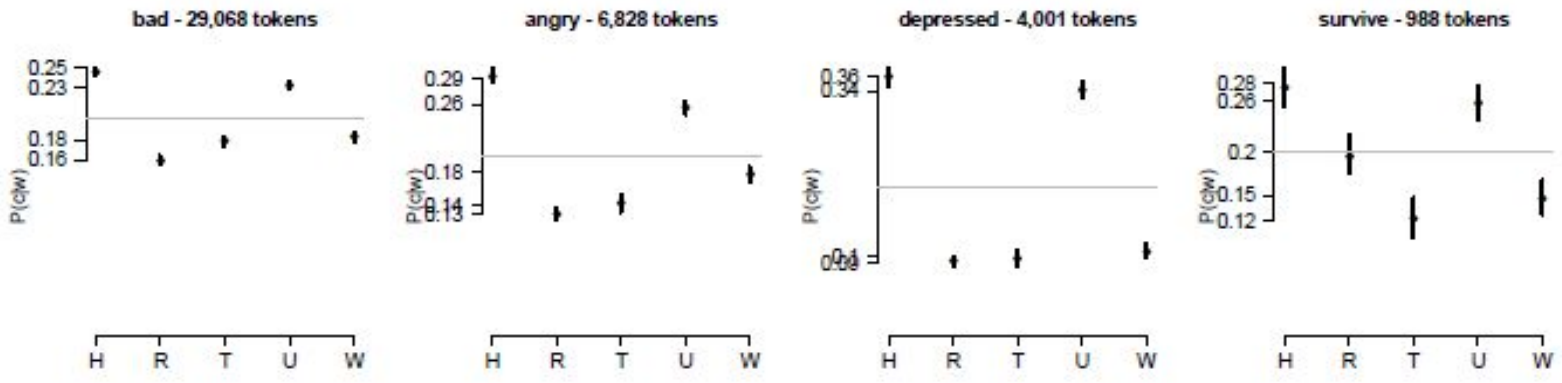


Figure: Words eliciting sympathetic reactions.

Acton and Potts, 2013; Potts, 2013

- Что же особенного в употреблении указательных местоимений у Palin?

- 16 интервью Palin на шоу
- + интервью до и после нее
- всего 48 интервью

Acton and Potts, 2013; Potts, 2013

Квантитативный анализ

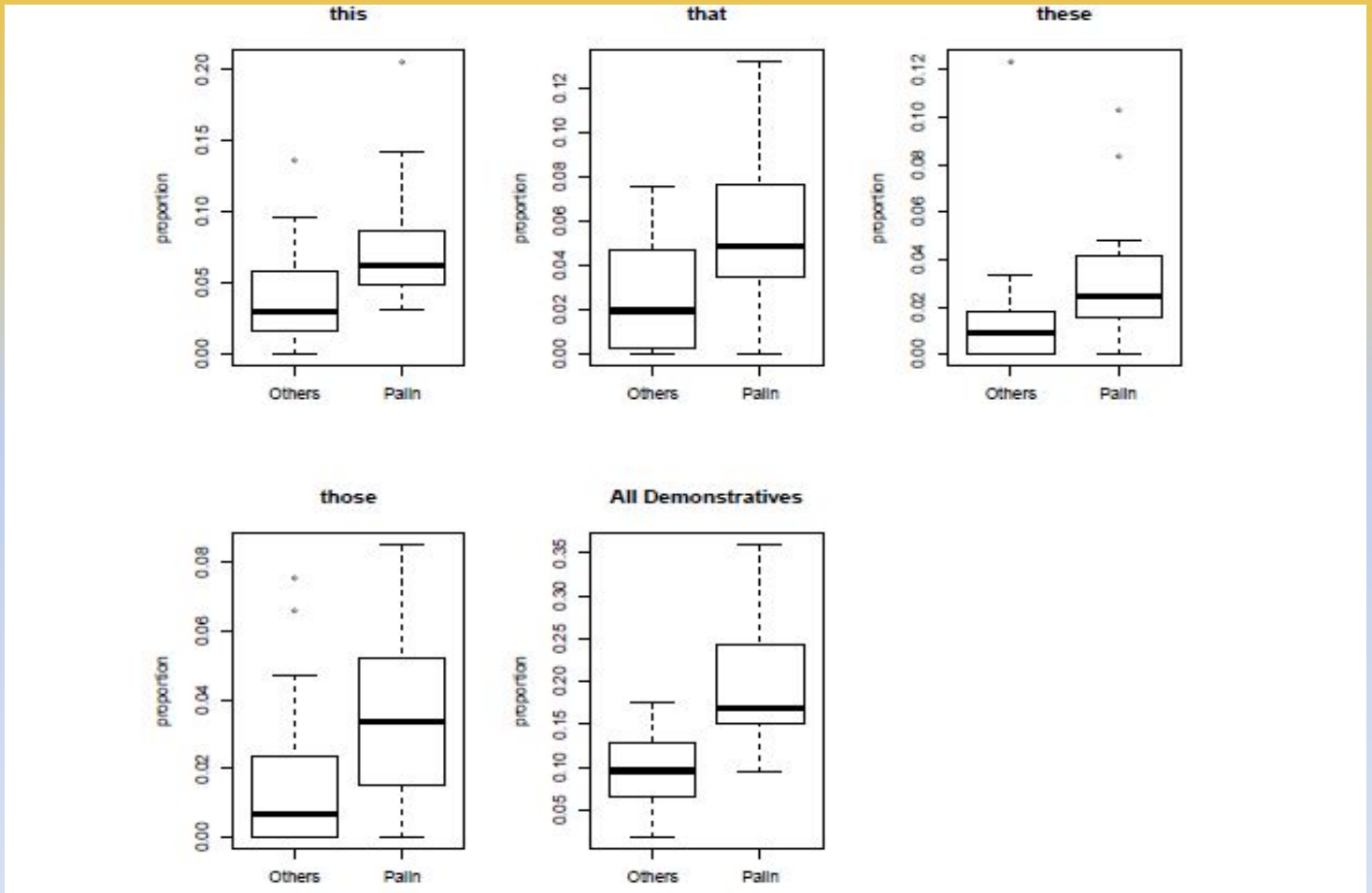


Figure: Non-pronominal dems. as a proportion of all determiners.

Acton and Potts, 2013; Potts, 2013

Полярные мнения – почему?

FoxNews.com:

- *“We feel like she talks like we do.”*
- *“She talked like real people to real people”*

Huffington Post:

- *“illusion of straight-talking”*
- *“pseudo-folksiness and fundamental dishonesty”*

Acton and Potts, 2013; Potts, 2013

Свойскость

Указательные местоимения в непрямом значении создают эффект **солидарности, свойскости** (термин восходит к лекциям Сандро Васильевича Кодзасова, мой перевод для solidarity из анализа статьи)

Проявления свойскости **уместны при совпадении отношения к сообщаемому, мнений**

При несовпадении эффект обратный: навязанной солидарности, **навязанной близости мнений, отношения к**

Acton and Potts, 2013; Potts, 2013

Квантитативный анализ

1. Группы с указательными местоимениями у Palin длиннее, более описательны, чем у других

2. Более провокационный контекст

"***these good, hard-working, average, everyday, patriotic Americans*** who want to see the *positive change* in our country that they *deserve*"

Acton and Potts, 2013; Potts, 2013

НКРЯ

- Какого цвета бывают эмоции?

НКРЯ

- Подсказка. Для того, чтобы задать словосочетания «цвет + эмоция», в лексико-грамматическом поиске, воспользовавшись полем для поиска по семантическим признакам, в качестве первого слова задайте прилагательное цвета, а второго — существительное, обозначающее эмоцию.

НКРЯ

- В текстах какой тематики в НКРЯ чаще всего встречаются слова, начинающиеся с квазиприставки «квази-», например, «квазипатриот»?

НКРЯ

- Подсказка. Поиск НКРЯ дает возможность задавать только часть искомого слова с помощью звездочки (*). Затем перейдите в раздел статистики по гиперссылке над выдачей и найдите данные о тематике текстов, вошедших в выдачу.

- В каком году в НКРЯ впервые встретилось слово, начинающееся с квазиприставки «псевдо-» («псевдовещь», но не «псевдоним»)?

- Подсказка. Поиск НКРЯ дает возможность задавать только часть искомого слова с помощью звездочки (*). Не забудьте исключить не интересующие вас слова с помощью знака минус (-). Для того, чтобы упорядочить результаты по году создания, поменяйте порядок сортировки во вкладке «настройки».

- Какой автор, представленный в корпусе, чаще всего использует «животные» прилагательные для характеристики эмоций?

- Подсказка. Следует искать словосочетания, состоящие из отыменного прилагательного, образованного от названия животного, и существительного, обозначающего некоторый вид эмоций. Для задания такого словосочетания воспользуйтесь поиском по семантическим характеристикам, после чего ознакомьтесь со статистикой по авторам.

- Сколько раз в НКРЯ встречается имя «Надежда»?

Подсказки

- Так как не во всем НКРЯ снята омонимия, можно попробовать оценить количество вхождений этого имени на основе того, сколько раз в НКРЯ встретилось слово «Надежда», написанное с заглавной буквы. Для этого в поле «доп. признаки» расширенного поиска укажите лексему «надежда», а в дополнительных признаках поставьте галочку напротив параметра «слово с заглавной буквы».
- Можно также добавить признак имя собственное (сем. признаки).
- Скорее всего, Вы всё равно столкнетесь с некоторым количеством ложноположительных результатов
- https://en.wikipedia.org/wiki/False_positives_and_false_negatives

- Какая квазиприставка — «экстра-» или «ультра-» — стала использоваться в прилагательных раньше?

- Подсказка. Поиск НКРЯ дает возможность задавать часть искомого слова с помощью звездочки (*). Составьте по запросу для каждой единицы и отсортируйте обе выдачи по дате.
- Хорошо бы исключить некоторые лексемы, например, ультрафиолетовый

- Сколько раз за 1990-2000 годы в текстах, созданных мужчинами, в корпусе упоминается одежда?

- Подсказка. Задайте подкорпус, состоящий из текстов, созданных мужчинами в указанный срок, после чего воспользуйтесь поиском по подкорпусу по семантическим признакам и найдите существительные-предметы одежды. Ответом на вопрос будет найденное количество вхождений.

Подведём итоги

- дескриптивный, а не прескриптивный подход;
- понятие нормы в лингвистике – что нам может поведать корпусная лингвистика?
- большие данные – будем учиться их анализировать статистически;
- сафари, а иногда и джунгли – нужна смекалка.

Литература

- Acton, Eric K. and Christopher Potts. 2014. That straight talk: Sarah Palin and the sociolinguistics of demonstratives. *Journal of Sociolinguistics* 18(1): 3-31.
- Christopher Potts [joint research with Eric Acton]. 2013. Cravin' that straight talk: the latent affective meaning of demonstratives. Workshop on Computational Social Sciences, Stanford, Jan 11.
- Marie-Catherine de Marneffe, Christopher D. Manning and Christopher Potts. Veridicality and utterance understanding. https://web.stanford.edu/~cgpotts/papers/Factbank_ICSC.pdf
- Копотев, Михаил. 2014. Введение в корпусную лингвистику. Главы 1—13.

Спасибо за внимание!