Чего ни хватишься, всё есть

Доказательства (не)существования в лингвистике

Александр Пиперски

[1]Кафедра компьютерной лингвистики Института лингвистики РГГУ

[2]Лаборатория социолингвистики ШАГИ РАНХиГС

16.07.2014

Что это было?

Слова из КРОНГАУЗ

- *грузок*
- *грузон*
- *конга*
- *нарок*
- *угарок*

Проблема

Засчитывать ли эти слова? $\Longleftrightarrow$

$\Longleftrightarrow$ **существуют ли они в русском языке?**

1. Постановка проблемы
2. Как доказать (не)существование слова?
   - Словари
   - Интернет
   - Корпуса
3. Выводы

1 Постановка проблемы

2 Как доказать (не)существование слова?
- Словари
- Интернет
- Корпуса

3 Выводы

Постановка проблемы

Является ли некоторая последовательность фонем / букв словом русского языка?

Зачем это нужно?

- Игры со словами
- Составление словарей
- Просто интересно (= глубокий научный смысл) :)

1 Постановка проблемы

2 Как доказать (не)существование слова?

- Словари
- Интернет
- Корпуса

3 Выводы

1 Постановка проблемы

2 Как доказать (не)существование слова?
- Словари
- Интернет
- Корпуса

3 Выводы

Словарь

Стандартный способ проверки существования слова — посмотреть в словаре

Проблемы

- Каким словарём пользоваться?
- Словари содержат странные слова
- Словари не содержат нужные слова

Реальный пример

Tony Augarde. 1992. *The Oxford A to Z of Word Games*. Oxford: Oxford University Press

Английские слова вида G.....l

The word *Gobi* was disqualified because it is not in the *Concise Oxford Dictionary* (chosen as the reference source before the game started); but *Guarani* was acceptable because it *is* in this dictionary (p. 5).

- Хорош ли такой метод?

Лакуны в словарях

- Систематические лакуны: географические названия, названия жителей и т. п. в русских толковых словарях
- Словари не успевают за новыми словами: *прокрастинация*, *селфи*, ...
- **$\Longrightarrow$ словари — плохой источник знания о (не)существовании слов**

1 Постановка проблемы

2 Как доказать (не)существование слова?
- Словари
- Интернет
- Корпуса

3 Выводы

Найдётся всё

- Девиз Яндекса
- Но верно и про Google

Случайные 5-буквенные последовательности

- *мёкьн*
- *пгщем*
- *мъпэд*
- *хыыид*
- *лйырё*
- *ъетщв*
- *ньрщд*
- *радтм*
- *миэоц*
- *ерсзч*

Что найдётся в Google?

Сколько существует случайных 5-буквенных последовательностей?

$N = 33^5 = 39135393$

- Существует примерно 20 млн 5-буквенных последовательностей, которые находятся в Google
- Сколько слов есть в русских словарях? Даль: ~200 тыс., Зализняк: ~100 тыс.

Ручная проверка

- Не хочется считать любые последовательности кириллических символов, которые встречаются в Google, русскими словами
- Необходимо хотя бы просмотреть вручную контексты
- Что находится на *ъетщв*, *радтм* и проч.?

Уточняем правила генерации

- Большинство последовательностей случайных букв противоречат правилам русской графики (*мъпэд*, *лйырё*, *ъетщв*) или фонотактики (*мёкьн*, *пгщем*, *хыыид*, *ньрщд*, *радтм*, *ерсзч*)
- Одно полунормальное слово: *миэоц*
- Нужны другие принципы генерации!

И снова КРОНГАУЗ

- Составляем 5-буквенные слова вида CVCVC из слова *КРОНГАУЗ*
  (C = согласный, V = гласный)

Сколько существует таких слов?

- Теоретически:
  $N = (5 \times 4 \times 3) \times (3 \times 2) = 360$
- А реально?

Составление слов по словарям

http://4maf.ru/anagram.php

- 9 слов вида CVCVC из 360 возможных
- *газок*, *газон*, *гуран*, *загон*, *закон*, *зарок*, *кагор*, *разок*, *розан*

10 случайных слов из 360

- *загун*
- *газор*
- *нарог*
- *рагун*
- *казог*
- *нукар*
- *зонак*
- *зунок*
- *горак*
- *гарон*

Сколько из них — русские слова?

Проверяем в Google

- Любое такое слово — имя собственное: Рагун — город в Германии, Мэтью Гарон — хоккеист, ...
- А есть ли всё-таки нарицательные?

10 случайных слов из 360

- *загун*
- *газор*
- *нарог*
- *рагун*
- *казог*
- *нукар*
- *зонак*
- *зунок*
- *горак*
- *гарон*

загун

- Термины российского архитектурного наследия. Плужников В.И., 1995
  загун — сайник
- Помощник кроссвордиста
  САЙНИК (=загун) — крытый загон для лошадей, в усадьбе коми-зырян

нарог

- Даль
  на́рог? м. *ниж.-мак.* стрела, пускаемая с луга || *Зап.* сошник, лемех
- *ниж.-мак.* = слово из Макарьевского уезда Нижегородской губернии
- Есть и в этимологическом словаре Фасмера

загун и *нарог*

- 2 из 10 $\Longrightarrow$ примерно 72 из 360
- NB: при автоматическом подборе слов по словарю — 9 из 360
- Но хотим ли мы считать *загун* и *нарог* словами русского языка?

Интернет: итог

- Кажется, что Интернет позволяет признать существующими слишком много слов
- Лучше использовать специальные подборки текстов — корпуса

1 Постановка проблемы

2 Как доказать (не)существование слова?
- Словари
- Интернет
- Корпуса

3 Выводы

Корпуса

- Корпус — собрание текстов, снабжённых лингвистической разметкой и системой поиска
- Какие вы знаете корпуса русского языка?
- NB: «корпус» — a corpus, а не the corpus

Корпуса

- Национальный корпус русского языка (НКРЯ): 230 млн слов
- Если слово встречается в НКРЯ 1 раз, есть ли оно в русском языке?

ormer-fidler.livejournal.com/22044.html (2006)

inna2: Спрашивает у меня тут нерусскоязычный коллега: "Какая разница между "проснулся" и "разбудился"? Я, естественно, отвечаю, что слова "разбудился" в русском языке вообще нет. Он: "А в интернете полно примеров". <...>

ormer-fidler: А, собственно, в чём проблема? Откуда такие панические интонации? Есть такой глагол, и у Даля, и в других словарях. И в корпусе есть как минимум один пример

Много или мало?

- Сейчас в НКРЯ не 1 пример на *разбудиться*, а 5 — но много это или мало?
- Надо учитывать объём корпуса
- $f = 5 \,/\, 230000000 \times 1000000 = 0,02$ вхождения на миллион слов

Много или мало?

- 0,02 вхождения на миллион слов — это много или мало?
- А какие слова не встречаются в НКРЯ ни разу?

0 раз в НКРЯ

- *прокрастинация*
- *селфи*
- *североморец*
- *клубнеобразование*
- *экономист-международник*
- *микруха*

Можно ли считать их несуществующими?

Ноль или не ноль?

- 0 или не 0? Зависит от размера корпуса
- НКРЯ (230 млн) vs. ruTenTen (16 млрд)

0 раз в НКРЯ — а в ruTenTen?

- *прокрастинация* : 966 (0,06 на млн)
- *селфи* : 6 (0,0004 на млн)
- *североморец* : 1142 (0,07 на млн)
- *клубнеобразование* : 505 (0,03 на млн)
- *экономист-международник* : 669 (0,04 на млн)
- *микруха* : 2875 (0,18 на млн)

Нули в большом корпусе

- А если 0 раз в ruTenTen?
- А если бы корпус был больше?

Корпус как выборка

- Корпус — не **генеральная совокупность**, а **выборка**
- По выборке нельзя сделать точных выводов о свойствах генеральной совокупности — только предположения разной степени достоверности

Аналогия: соцопрос

- «Понравилась ли вам эта лекция»?
- Опрошено 20 из 100 участников ЛЛШ
- 8 участникам ЛЛШ из 20 лекция понравилась
- Скольким из 100 участников ЛЛШ понравилась лекция?

Аналогия: соцопрос

- Может быть, 40 из 100
- Не исключено, что 30 из 100 или 50 из 100
- Маловероятно, что 10 из 100 или 80 из 100
- 95%-ный **доверительный интервал**: от 20 до 64 из 100

Вопрос на внимательность

В какие моменты лекции я пренебрёг идеей 95%-ного доверительного интервала?

2 из 10 — не 20%, а от 4% до 56%

5 из 10 — не 50%, а от 24% до 76%

Аналогия: соцопрос

- А если лекция понравилась 0 из 20?
- 95%-ный доверительный интервал: от 0 до 20 из 100
- NB: есть разные методы подсчёта доверительных интервалов (Newcombe 1998)

Доверительные интервалы и корпуса

- Слово X: 0 раз в НКРЯ (230 млн)
- 95%-ный доверительный интервал:
  от 0 до 5 из 230 млн (от 0 до 0,02 на млн)

Доверительные интервалы и корпуса

- Общее правило:
- Если слово встретилось в корпусе 0 раз, то оно с 95%-ной вероятностью должно встретиться в корпусе такого объёма и такого состава от 0 до 5 раз
- (статистические подробности — в кулуарах)

Корпуса и доказательства несуществования

- Ни про какое слово нельзя достоверно сказать, что его нет, даже если оно не встретилось в корпусе
- $\Longrightarrow$ доказательств несуществования в лингвистике не существует!

Корпуса и доказательства существования

- Какой частотности в корпусе достаточно, чтобы считать слово существующим в той разновидности языка, которую этот корпус представляет?
- Мой ответ: **не знаю**
- $\Longrightarrow$ доказательств существования в лингвистике тоже не существует!

1. Постановка проблемы
2. Как доказать (не)существование слова?
   - Словари
   - Интернет
   - Корпуса
3. Выводы

Выводы

- Словари и Интернет имеют сильные недостатки, если надо ответить на вопрос о (не)существовании слова
- Корпуса лучше, но надо правильно ими пользоваться

Выводы

- Ни существование, ни несуществование в лингвистике нельзя доказать, а значит, бинарное противопоставление «существует» / «не существует» бессмысленно

Спасибо за внимание!