Способы оценки корпусов

Е. Клячко

Оплинг

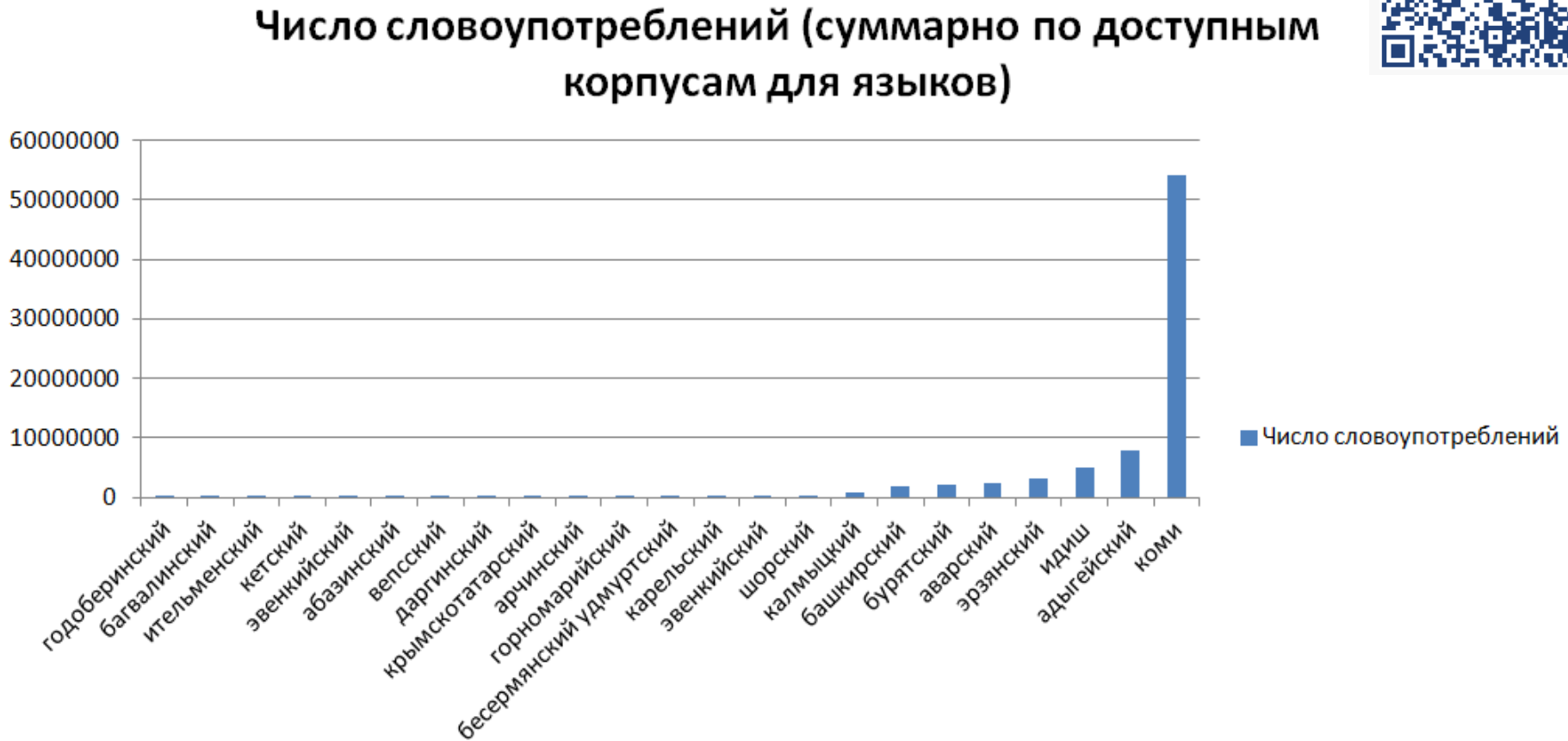
Корпус устных полевых записей: сомнения

- мало данных — насколько имеем право применять статистические методы
- несбалансированность
- большая вариативность

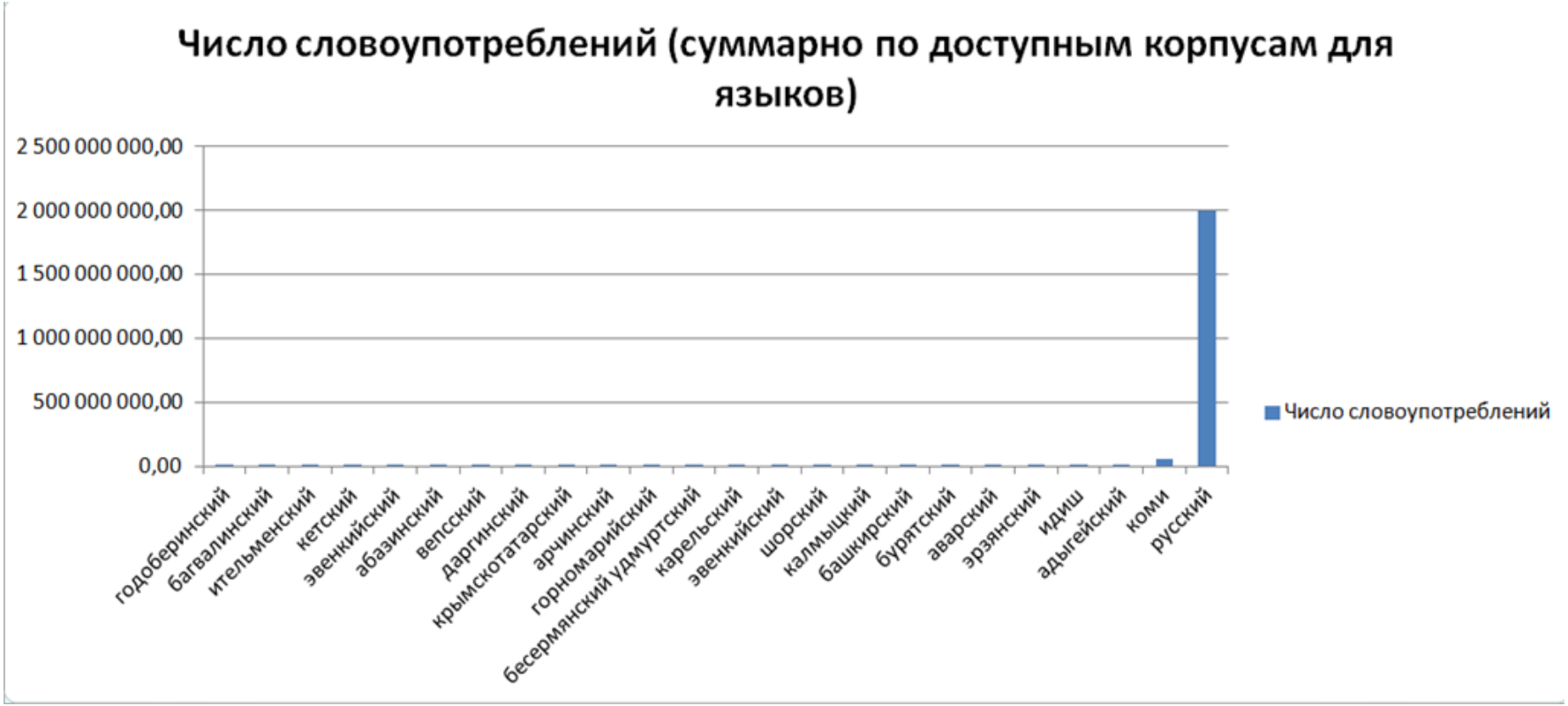
⇒ Есть ли способ формализовать эти проблемы?

и разрешить сомнения?

Объем корпусов текстов на языках России



Объем корпусов текстов на языках России



Критерии “пригодности”

- для лингвистических исследований
- для машинной обработки

Проблемы: вариативность

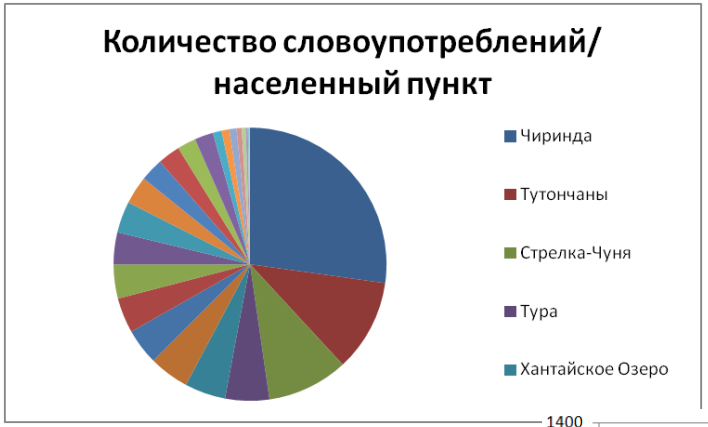
- устная речь: стандартизация?
- расхождения в разметке
  - изменения к подходу в глоссировании
  - “омонимичные” показатели
  - ошибки человека
  - ошибки автоматической разметки

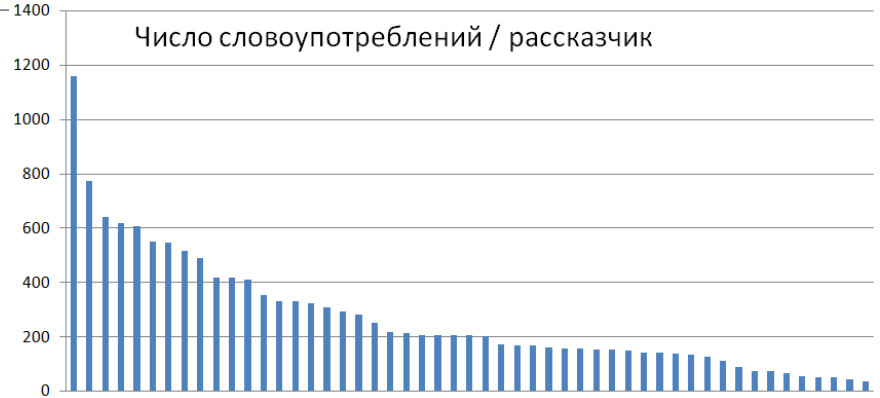
⇒ инструменты поиска “подозрительных” разборов

Репрезентативность и сбалансированность

- число словоупотреблений
- жанры и тематика
  - фольклор, этнография, воспоминания
- монолог vs диалог vs полилог
- выбор рассказчиков / текстов для включения в корпус
  - ограниченность ресурсов
  - гендер / возраст / профессия / …
  - “хороший рассказчик”
  - предпочтения разметчика
- выбор при самозаписи
- представленность диалектов

Пример эвенкийского корпуса: https://minlang.iling-ran.ru/corpora/evenki





Литература?

- автоматизированная чистка наборов данных:
  - поиск ошибо
  - outlier detection — выявление аномалий
- ручная оценка качества
- энтропия как оценка сбалансированности?
- рекомендации по статистическим методам:
  - Egbert J., Larsson T., Biber D. Doing linguistics with a corpus: Methodological considerations for the everyday user. – Cambridge University Press, 2020.
  - Paquot M., Gries S. T. (ed.). A practical handbook of corpus linguistics. – Springer Nature, 2021.

Интерпретируемо ли то, что мы можем получить по корпусу?

- частотные списки
- коллокации
- языковые модели

коллокация (по словоформам, порог=10)	перевод
a to	‘а то’ (*рус.*)
tar kiŋgit	‘тот Кингит’ (*персонаж сказки*)
tak i	‘так и’ (*рус.*)
tar ahī	‘та женщина’
tar bəjə	‘тот мужчина’
tar ələ	*дискурсивный маркер* (‘так вот…’)
taduk hələ	*дискурсивный маркер* (‘потом вот…’)
i wot	‘и вот’ (*рус.*)
taduk nuŋan	‘потом он(а/о)’

коллокация (по леммам, порог=10)	перевод
tuŋ ńikə-	так делать
d’u d’u-	дом дом
toɣo hī-	огонь дуть
bira ďapka	река берег
a to	а то (*рус.*)
mō mō-	дерево дерево
ə hula-	INTJ оставить
tar kiŋgit	тот Кингит
qun- təpkə-	сказать кричать

Языковые модели, вектора (word embeddings), опыт с TTS

- https://www.tensorflow.org/tutorials/text/word2vec?hl=en
- https://huggingface.co/blog/how-to-train
- The final training corpus has a size of 3 GB, which is still small – for your model, you will get better results the more data you can get to pretrain on.

Рекомендации по статистическим подходам (Egbert et al. 2020)

- пример, когда “проверка гипотезы” подводит
- использование conditional inference tree (деревья условного вывода)
  - интерпретируемая методика
  - меньше требований к объему и составу данных
- “возвращаться” от статистики к данным

Результаты

- публиковать подробные метаданные
- балансировать состав корпуса (?)
- автоматически отлавливать ошибки
- анализировать частотные списки