Полевые данные и компьютерные инструменты

Г. Мороз
Международная лаборатория языковой конвергенции, НИУ ВШЭ, Москва

Иркутский государственный университет
14 – 16 ноября 2022 г.
«Цифра» в социально-гуманитарных исследованиях: метод, поле, реальность?

План презентации

Малые языки в большой лингвистике

Устные корпуса

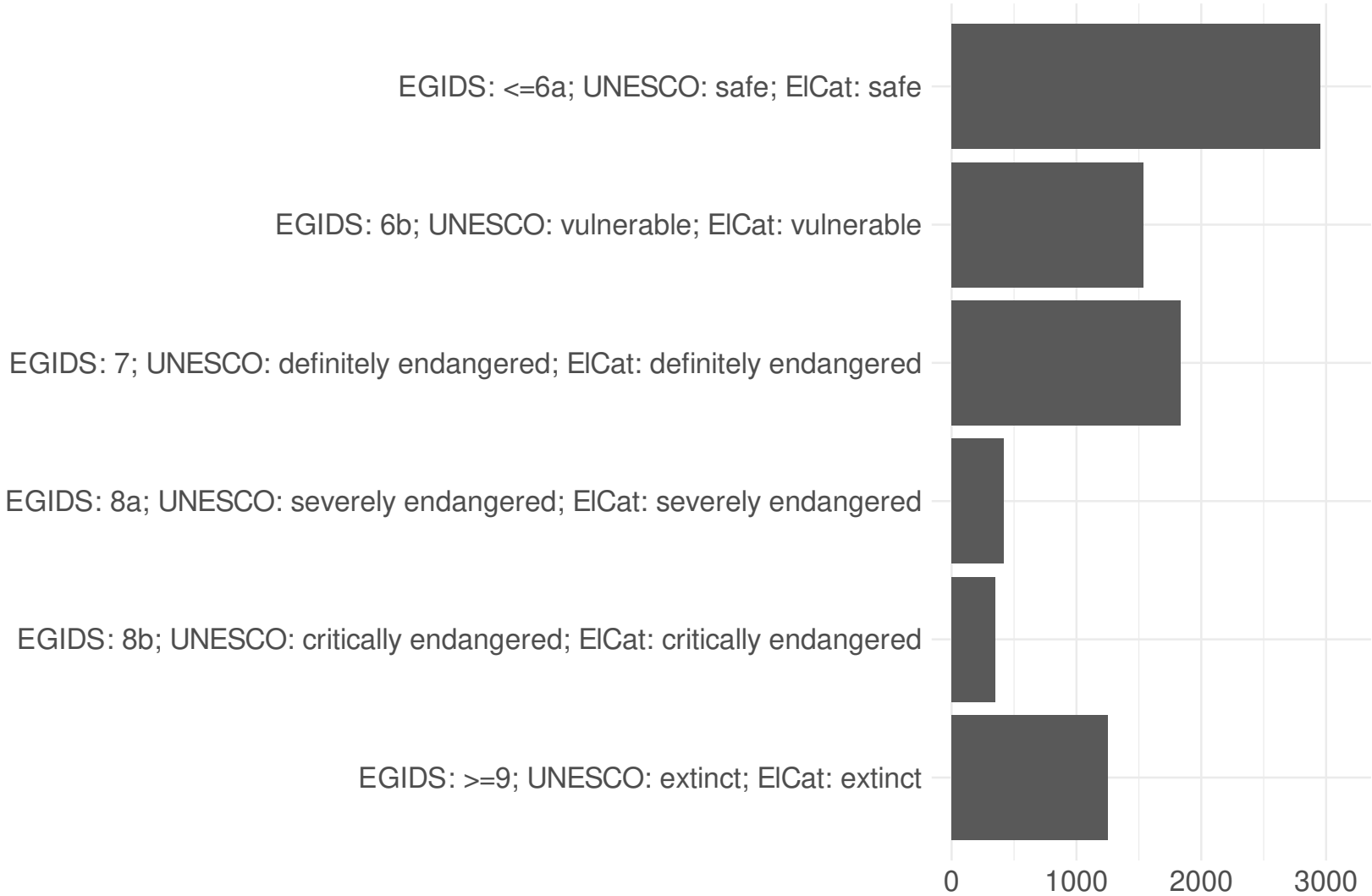
Морфологический анализаторы

Синтаксические парсеры

Какие бывают языки с точки зрения компьютерных инструментов?

- огромные (шкала EGIDS (Lewis and Simons 2010) = 1)
  - доступно очень много текстовых данных, сформировалась литература, есть здоровый интернет
  - доступны даже исторические данные, скажем, на целый век
- средние (шкала EGIDS от 2 до 5)
  - додступны грамматические описания, двуязычные словари
  - литературная традиция часто ограничена одним веком, и обычно немногочисленна
- малые (шкала EGIDS больше 5)
  - острая нехватка материалов

Какие бывают языки с точки зрения компьютерных инструментов?



на основе Glottolog v. 4.6.1

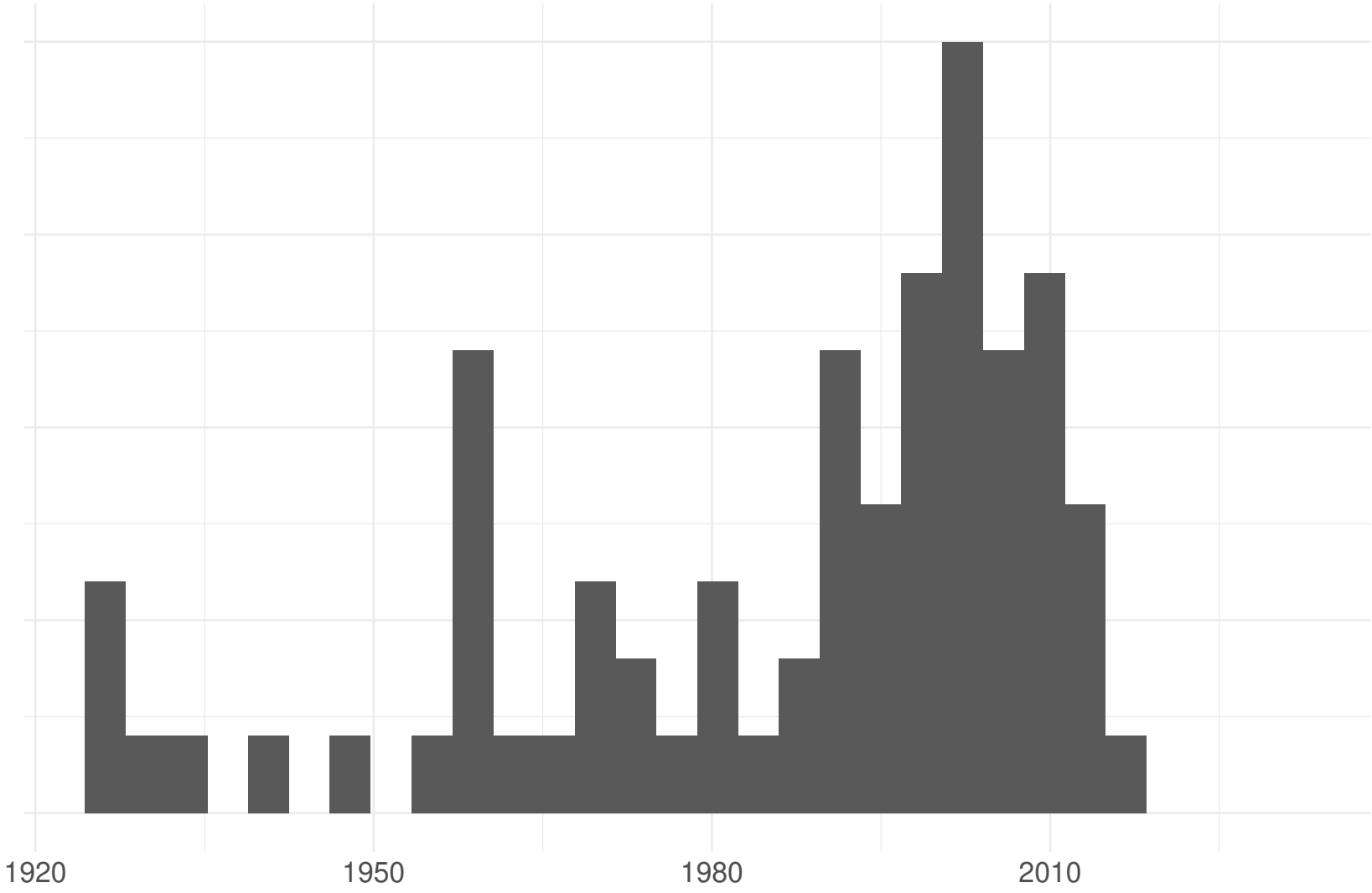
Как это влияет на инструменты компьютерной лингвистики для малых языков?

- нет данных или их очень мало, так что все что строится на нейросетях не работает
  - морфологический парсер?
  - синтаксический парсер?
  - распознавание/синтез речи?

Как это влияет на инструменты компьютерной лингвистики для малых языков?

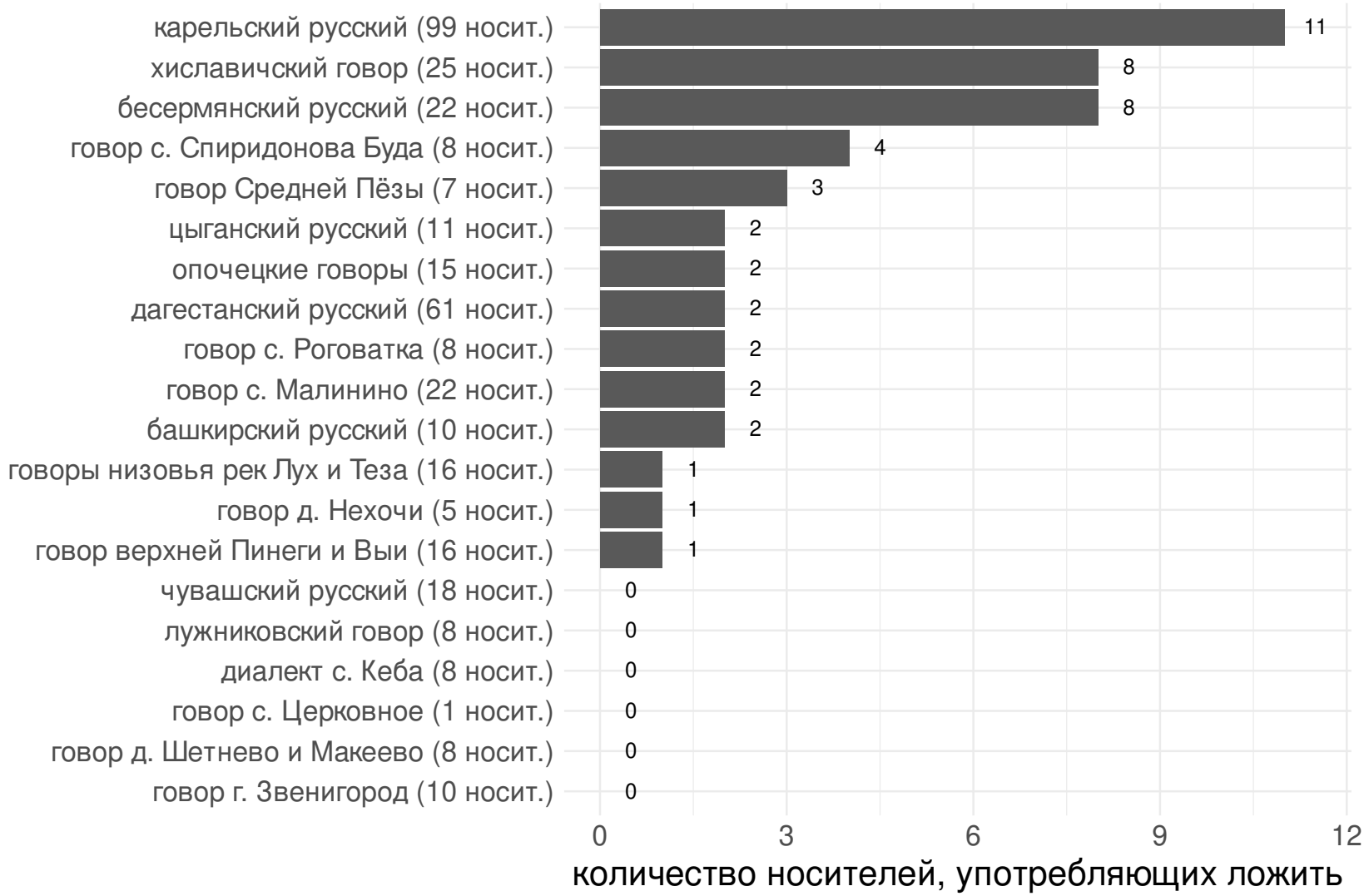
- нет данных или их очень мало, так что все что строится на нейросетях не работает
  - морфологический парсер?
  - синтаксический парсер?
  - распознавание/синтез речи?
- как следствие, не всегда можно вслепую копировать методы разработанные для больших языков

Лексема *ложить* в НКРЯ



77 уникальных авторов отфильтрованных из 141 примера

Лексема *ложить* в устных корусах



на основе устных корпусов
междунароной лаборатории языковой конвергенции

Как это влияет на инструменты компьютерной лингвистики для малых языков?

- нет данных или их очень мало, так что все что строится на нейросетях не работает
  - морфологический парсер?
  - синтаксический парсер?
  - распознавание/синтез речи?
- как следствие, не всегда можно вслепую копировать методы разработанные для больших языков

В результате, для малых языков необходимы особые инструменты и методы, которые бы позволяли облегчать или делать возможным лингвистический анализ и инструменты для языкового комьюнити (спеллчекеры, предективный набор и т. п.).

Инструменты для полевых лингвистов

- устные корпуса
  - без морфологической разметки
  - с морфологической разметкой
- морфологические анализаторы (трансдьюссеры lexd и twol)
- синтаксические парсеры (проект Universal Dependencies)

План презентации

Малые языки в большой лингвистике

Устные корпуса

Морфологический анализаторы

Синтаксические парсеры

План презентации

Малые языки в большой лингвистике

Устные корпуса

Морфологический анализаторы

Синтаксические парсеры

План презентации

Малые языки в большой лингвистике

Устные корпуса

Морфологический анализаторы

Синтаксические парсеры

References

M. P. Lewis and G. F. Simons. Assessing endangerment: expanding fishman's gids. 2010.