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

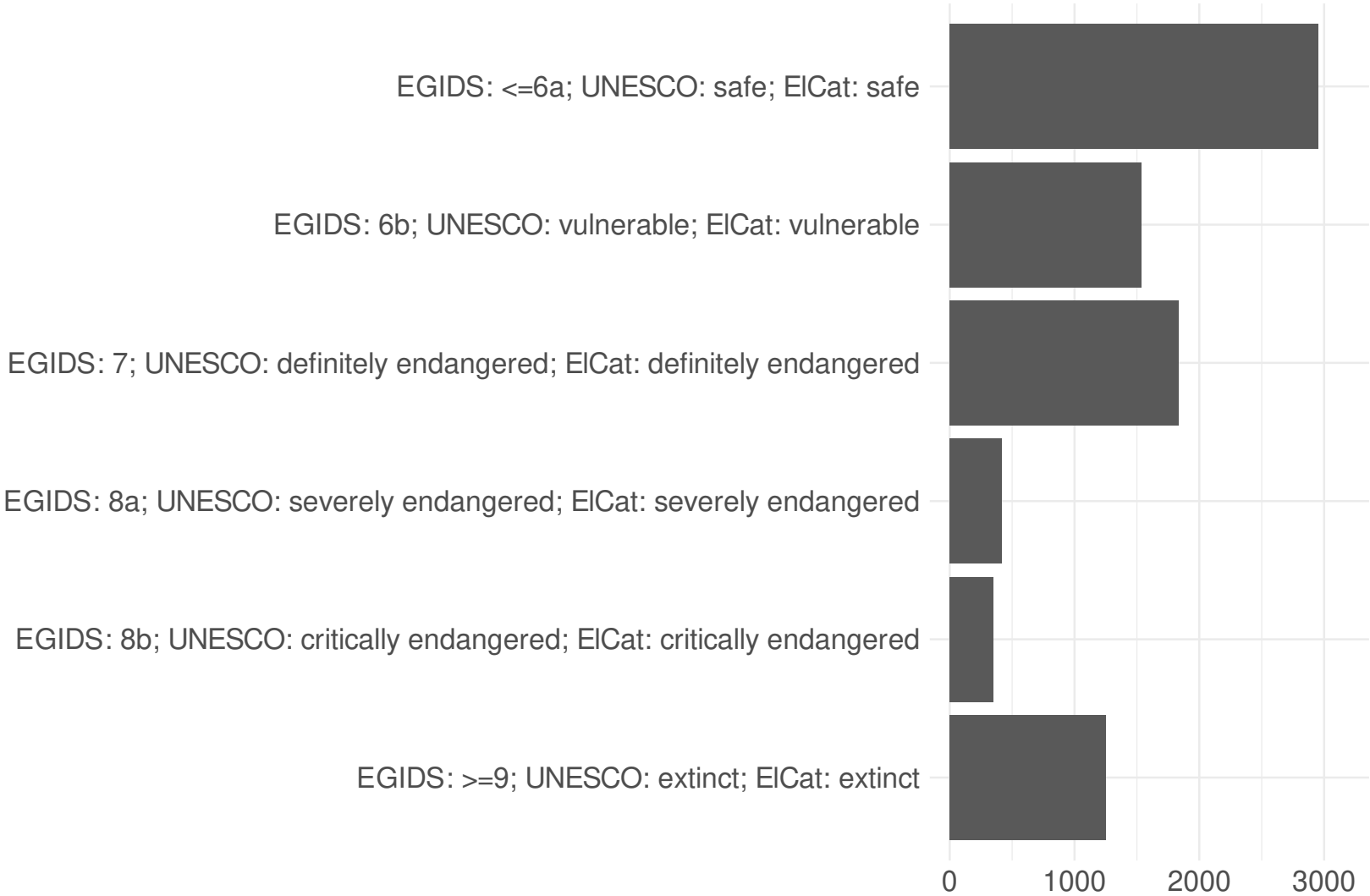
Устные корпуса

Морфологический анализаторы

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

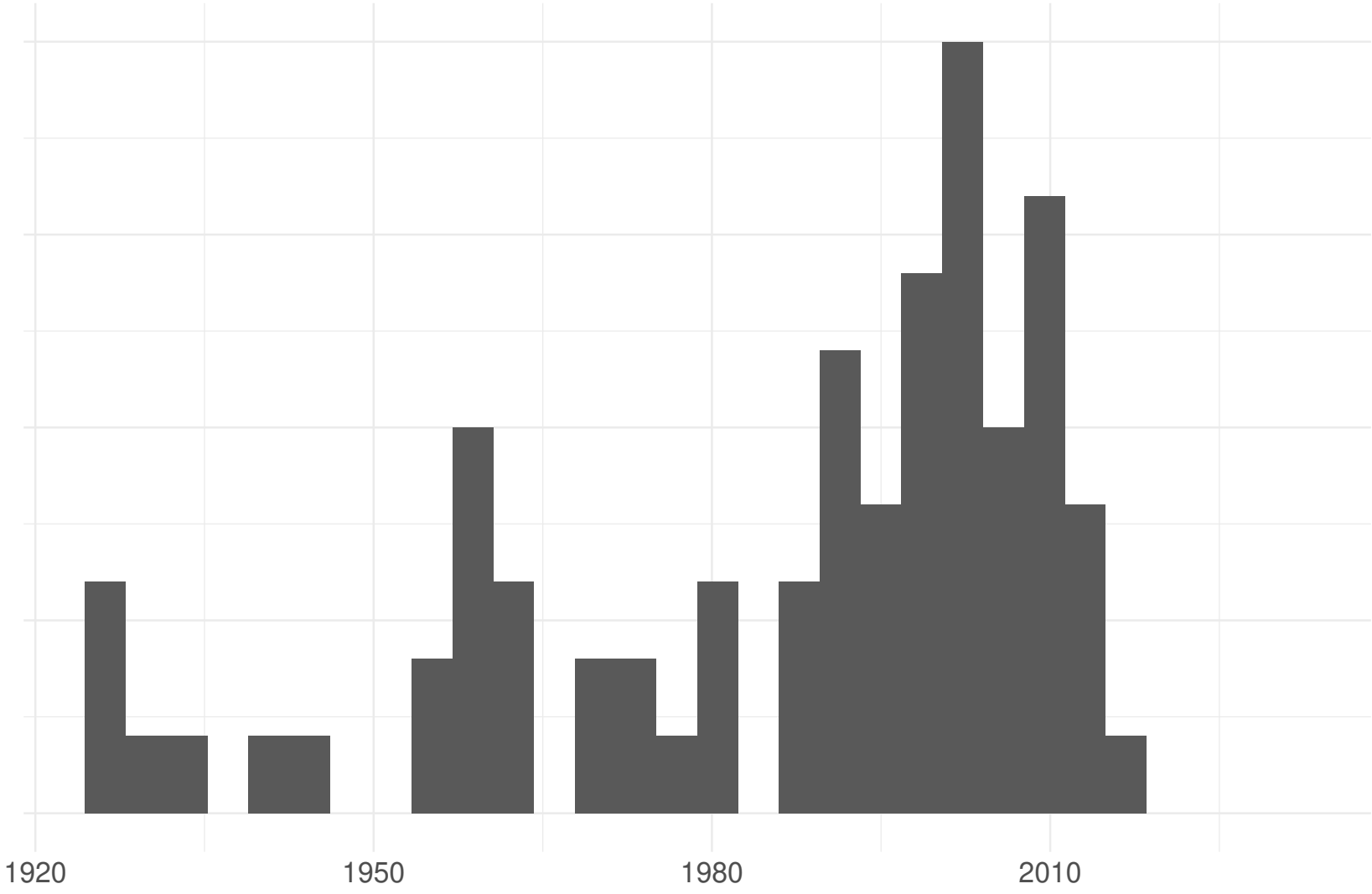
Как это влияет на инструменты компьютерной лингвистики для малых языков?

- 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?

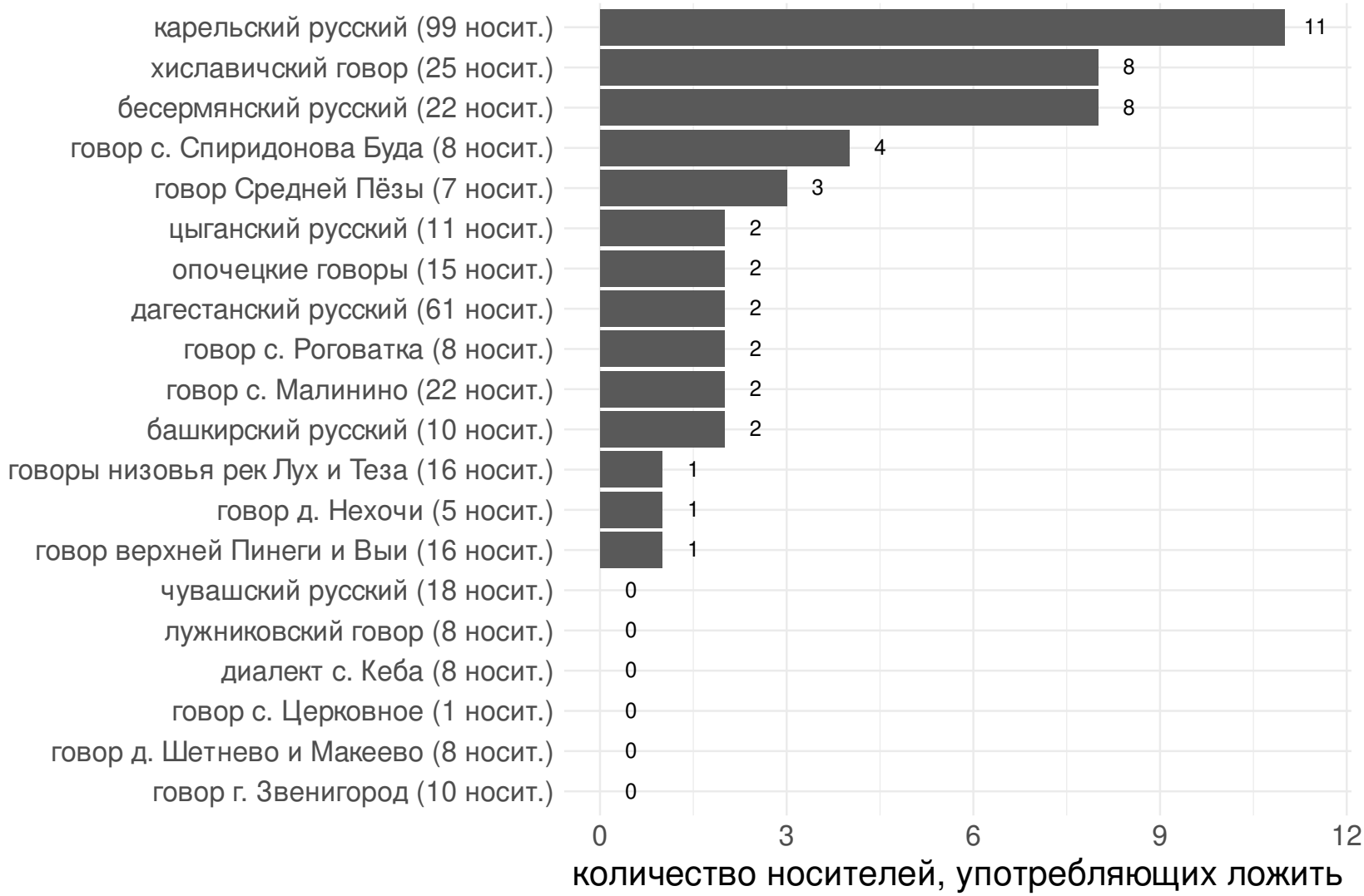
- нет данных или их очень мало, так что все что строится на нейросетях не работает
  - морфологический парсер?
  - синтаксический парсер?
  - распознавание/синтез речи?
- как следствие, не всегда можно вслепую копировать методы разработанные для больших языков

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

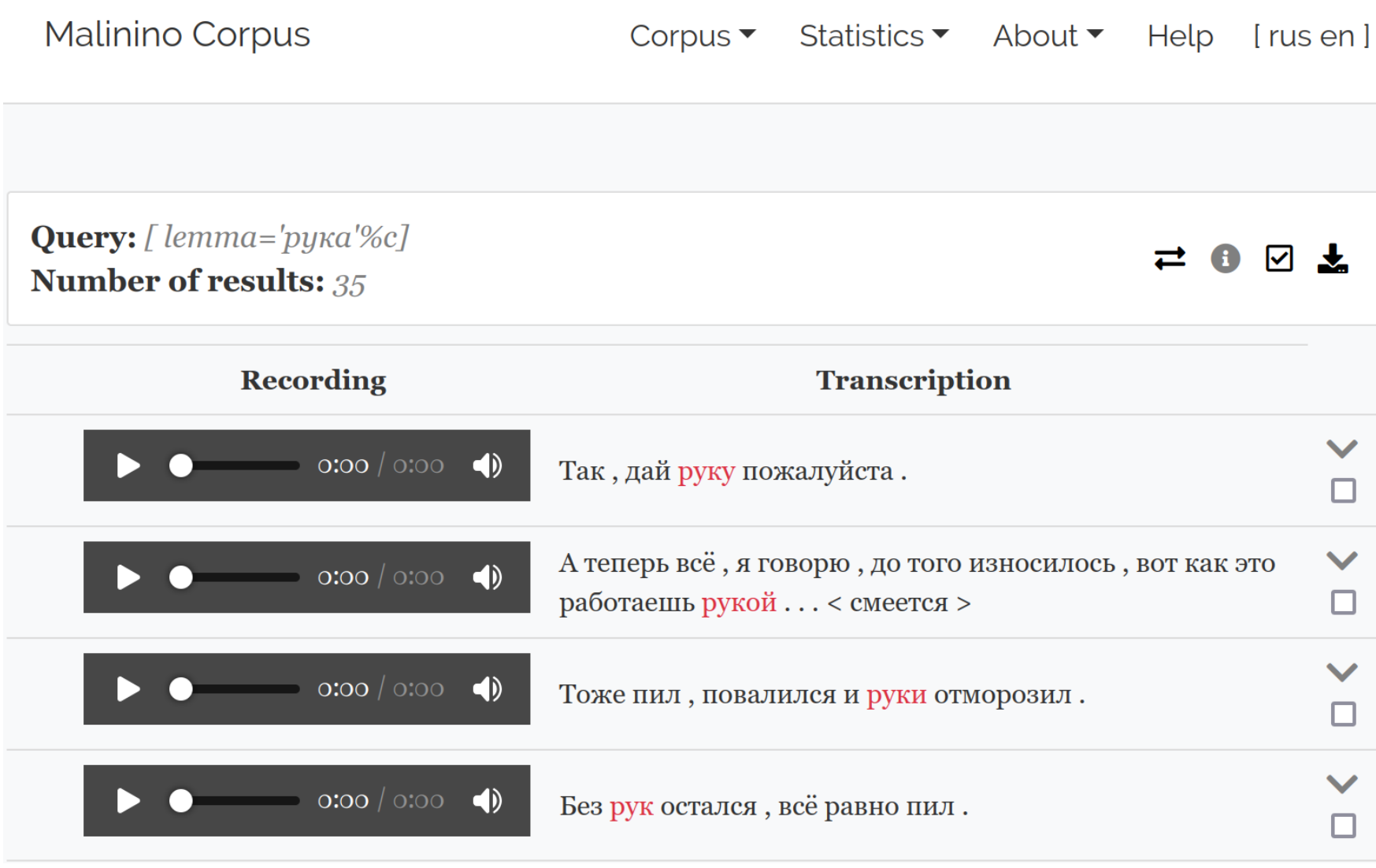
SpoCo (von Waldenfels and Woźniak 2016)

- устьянские говоры (959 782 сл.-у.);
- хиславичский говор (260 793 сл.-у.);
- говоры низовья рек Лух и Теза (146 350 сл.-у.);
- говор с. Малинино (138 943 сл.-у.);
- говор с. Роговатка (100 047 сл.-у.);
- говор д. Нехочи (88 965 сл.-у.);
- говор Средней Пёзы (79 566 сл.-у.);
- говор верхней Пинеги и Выи (70 803 сл.-у.);
- говор с. Спиридонова Буда (70 565 сл.-у.);
- опочецкие говоры (68 741 словоупотребление);
- лужниковский говор (68 666 сл.-у.);
- говор г. Звенигород (68 324 сл.-у.);
- говор д. Шетнево и Макеево (58 003 сл.-у.);
- говор с. Кеба (54 535 сл.-у.);
- говор с. Церковное (19 960 сл.-у.).

SpoCo (von Waldenfels and Woźniak 2016)

- дагестанский русский (227 885 сл.-у.);
- бесермянский русский (97 216 сл.-у.);
- башкирский русский (93 127 сл.-у.);
- карельский русский (74 014 сл.-у.);
- чувашский русский (46 307 сл.-у.);
- цыганский русский (41 767 сл.-у.);
- якутский русский (15 139 сл.-у.).

SpoCo (von Waldenfels and Woźniak 2016)



SpoCo (von Waldenfels and Woźniak 2016)

201008e_mgd

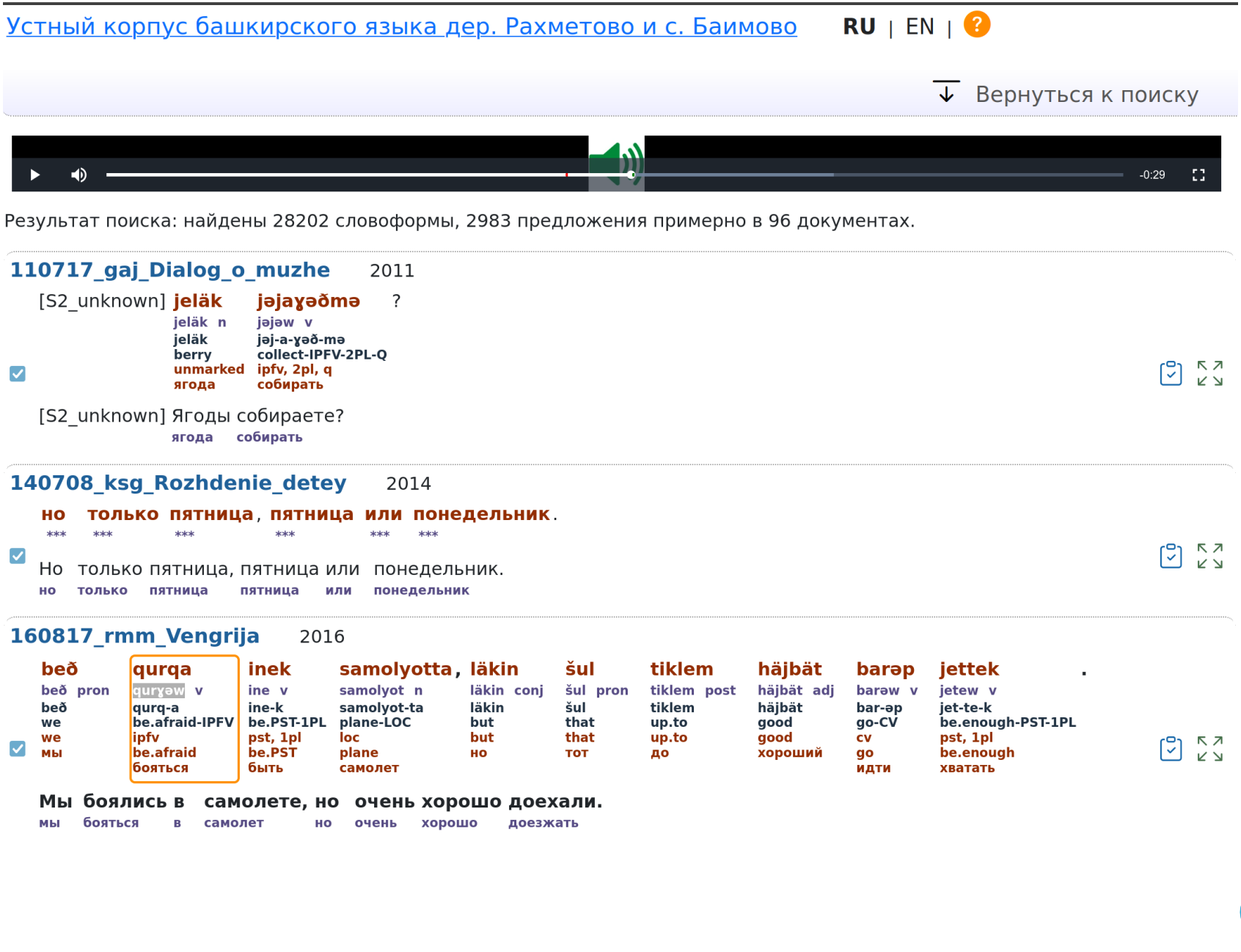
Слушать в одном файле (mp3)
First transcription: 0:0:0 Last transcription: 0:18:10

Interviewer: <>[i] ♩
mgd: <Рисовала>... Русалка.[i] ♩
Interviewer: Ага.[i] ♩ А масленницу делали у вас?[i] ♩
mgd: А?[i] ♩
Interviewer: А на масленницу не было?[i] ♩
mgd: О, а как же! И масленица бывает.[i] ♩ Тогда, бывало, катался с горы,[i] ♩
Interviewer: С горы?[i] ♩

Tsakorpus (Arkhangelskiy 2019)

- Абазинский
- Адыгейский
- Башкирский
- Кабардинский
- Хакасский
- Луговой Марийский

Tsakorpus (Arkhangelskiy 2019)



Все наши корпуса доступны здесь

http://lingconlab.ru/

План презентации

Малые языки в большой лингвистике

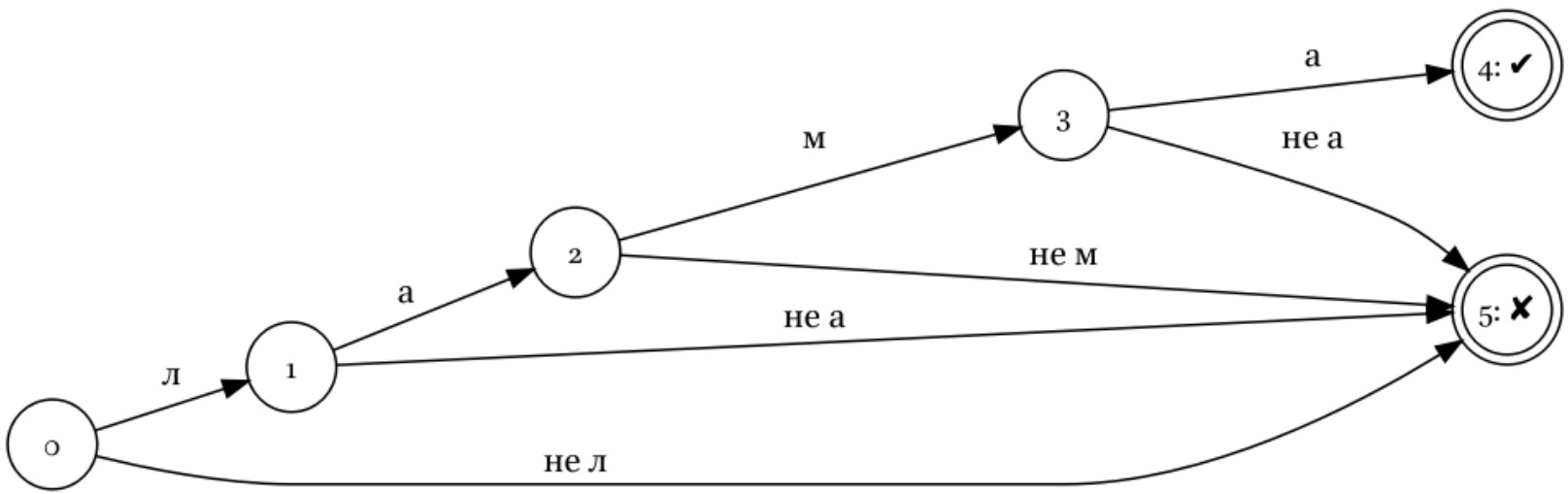
Устные корпуса

Морфологический анализаторы

Что такое трансдьюсер?

Трансдьюсер (конечный автомат с выходом) — это вид конечного автомата с двумя лентами памяти.

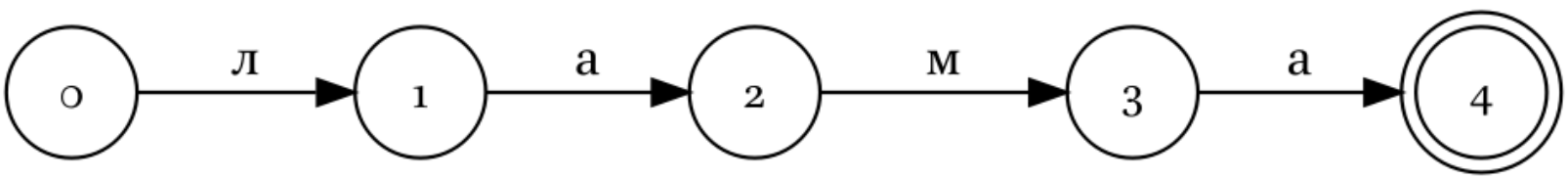
- Вот пример конечного автомата с одной лентой памяти. Он проверяте является поданное на вход слово словом *лама*:



Что такое трансдьюсер?

Трансдьюсер (конечный автомат с выходом) — это вид конечного автомата с двумя лентами памяти.

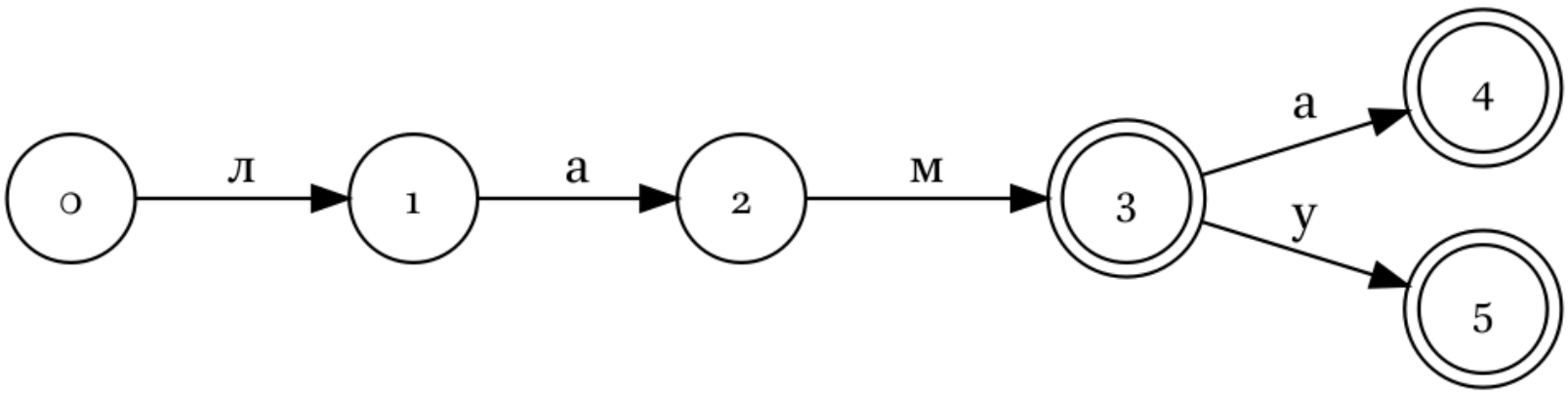
- Вот пример конечного автомата с одной лентой памяти. Он проверяте является поданное на вход слово словом *лама*.
- Обычно эти ветви “не X” не пишут:



Что такое трансдьюсер?

Трансдьюсер (конечный автомат с выходом) — это вид конечного автомата с двумя лентами памяти.

- Вот пример конечного автомата с одной лентой памяти. Он проверяте является поданное на вход слово словом *лама*.
- Обычно эти ветви “не Х” не пишут.
- Можно закодировать больше слов (*лама*, *ламу*, *лам*):



Что такое трансдьюсер?

У трансдьюсеров две ленты памяти, что можно воспринимать как переписывание:

- лама переходит в лама<n><nom><sg>
- ламу переходит в лама<n><acc><sg>
- лам переходит в лама<n><acc><pl>
- все остальное – ошибка

Зачем использовать трансдьюсер?

- они обратимы, так что анализ (ламу → лама<acc><sg>)и генерация (лама<acc><sg> → ламу) могут быть сделаны одним и тем же трансдьюсером

Зачем использовать трансдьюсер?

- они обратимы, так что анализ (`ламу → лама<acc><sg>`)и генерация (`лама<acc><sg> → ламу`) могут быть сделаны одним и тем же трансдьюсером
- их можно оптимизировать для быстрого поиска

Зачем использовать трансдьюсер?

- они обратимы, так что анализ (ламу → лама<acc><sg>)и генерация (лама<acc><sg> → ламу) могут быть сделаны одним и тем же трансдьюсером
- их можно оптимизировать для быстрого поиска
- их можно соединять с другими трансдьюсерами (например, транслитерация или даже перевод)

Зачем использовать трансдьюсер?

- они обратимы, так что анализ (ламу → лама<acc><sg>)и генерация (лама<acc><sg> → ламу) могут быть сделаны одним и тем же трансдьюсером
- их можно оптимизировать для быстрого поиска
- их можно соединять с другими трансдьюсерами (например, транслитерация или даже перевод)
- формализм трансдьюсеров позволяет описывать языковой материал в приближенном к лингвистическому описанию виде

Как использовать трансдьюсер?

- можно почитать (Beesley and Karttunen 2003; Karttunen and Beesley 1992)
- `lexd` — компилятор для морфологии (Swanson and Howell 2021)
- `twol` — компилятор для морфонологии

lexd пример (зиловский андийский)

```
PATTERNS
Numerals NumearalMarker

LEXICON Numerals
иʰшду                  # пять; five
ойлIи                   # шесть; six
гьокьу                  # семь; seven
бейкьи                  # восемь; eight
гьочIо                  # девять; nine

LEXICON NumearalMarker
<num>:гу
```

lexd пример (зиловский андийский)

```
PATTERNS
Numerals NumearalMarker

LEXICON Numerals
иʰшду                  # пять; five
ойлIи                   # шесть; six
гьокьу                  # семь; seven
бейкьи                  # восемь; eight
гьочIо                  # девять; nine

LEXICON NumearalMarker
<num>:гу

иʰшдугу:иʰшду<num>
ойлIигу:ойлIи<num>
гьокьугу:гьокьу<num>
бейкьигу:бейкьи<num>
гьочIогу:гьочIо<num>
```

Как разработать морфологический трансдьюсер?

- опишите морфологию и морфонологию используя доступные ресурсы

Как разработать морфологический трансдьюсер?

- опишите морфологию и морфонологию используя доступные ресурсы
- составьте словарь с словоизменительной аннотацией

Как разработать морфологический трансдьюсер?

- опишите морфологию и морфонологию используя доступные ресурсы
- составьте словарь с словоизменительной аннотацией
- опционально можно сделать список тестовых форм, которые трансдьюсер должен разбирать

Как разработать морфологический трансдьюсер?

- опишите морфологию и морфонологию используя доступные ресурсы
- составьте словарь с словоизменительной аннотацией
- опционально можно сделать список тестовых форм, которые трансдьюсер должен разбирать
- `чIe<NUM><num><obl.m><epent.m><an.sg><aff>` `чIегушубо` (Zilo Andi)
- проверьте ваш трансдьюсер на аннотированном (или даже неаннотированном) корпусе

Спасибо за внимание!

agricolamz@gmail.com

References

T. Arkhangelskiy. Corpora of social media in minority Uralic languages. In *Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages*. Association for Computational Linguistics, Tartu, Estonia, 2019.

K. R. Beesley and L. Karttunen. *Finite-state morphology: Xerox tools and techniques*. CSLI, Stanford, 2003.

L. Karttunen and K. R. Beesley. *Two-level rule compiler*. Xerox Corporation, Palo Alto Research Center, 1992.

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

D. Swanson and N. Howell. Lexd: a Finite-State lexicon compiler for non-suffixational morphologies. In M. Hämäläinen, N. Partanen, and K. Alnajjar, editors, *Multilingual facilitation*. 2021.

References

Ruprecht von Waldenfels and Michał Woźniak. Spoco-a simple and adaptable web interface for dialect corpora. *Journal for language technology and computational linguistics*, 31:155–170, 2016.