lingtypology: R пакет для лингвистического картографирования

Г. Мороз
Лаборатория языковой конвергенции, НИУ ВШЭ

18 декабря 2018 г.

Открытые лекции — «Городские данные»
Софт Культуры и Инфокультуры

ссылка на презентацию: tinyurl.com/ycx46od6

#тыжлингвист

#тыжлингвист

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух

ссылка на презентацию: tinyurl.com/ycx46od6

#тыжлингвист

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова

ссылка на презентацию: tinyurl.com/ycx46od6

#тыжлингвист

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова
- пишет без ошибок и знает все правила орфографии

ссылка на презентацию: tinyurl.com/ycx46od6

#тыжлингвист

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова
- пишет без ошибок и знает все правила орфографии
- не знает математики и программирования

ссылка на презентацию: tinyurl.com/ycx46od6

#тыжлингвист

- умеет читать на всех письменностях мира
- знает все языки на свете
- умеет распознавать каждый язык на слух
- может рассказать о происхождении каждого слова
- пишет без ошибок и знает все правила орфографии
- не знает математики и программирования

все вышеперечисленное, конечно, неправда

ссылка на презентацию: tinyurl.com/ycx46od6

Лингвистика

- прескриптивная

Лингвистика

- прескриптивная
- вся остальная
  - исследования грамматики языка и языкового разнообразия
  - исследования распределения грамматических особенностей в языках мира
  - исследования когнитивных способностей человека и других животных, связанных с языком
  - исследования в области NLP и их приложения
  - исследования в области синтеза и распознования речи и языка
  - создание компьютерных инструментов для решения самых разных задач

Еще бывает *компьютерная лингвистика*:

- вспомогательные инструменты лингвистического исследования и документации
- Computational linguistics
- NLP

ссылка на презентацию: tinyurl.com/ycx46od6

лингвистические базы данных

Лингвистические базы данных

Корпуса — базы данных языкового материала

- корпус литературных текстов
- Русский национальный корпус
- аудио и видео корпуса

ссылка на презентацию: tinyurl.com/ycx46od6

Лингвистические базы данных

Корпуса — базы данных языкового материала

- корпус литературных текстов
- Русский национальный корпус
- аудио и видео корпуса
  - настоящая речь, а не тексты
  - см., например, Корпус бассейна реки Устья
  - см., например, Корпус села Роговатка
  - см., например, корпус русского жестового языка
  - см., например, Уошо

Базы данных языковых структуру

- The World Atlas of Language Structures (WALS)
- World Atlas of Varieties of English (eWAVE)
- Glottolog — система ссылок на языки мира.
- ... и другие

ссылка на презентацию: tinyurl.com/ycx46od6

Зачем лингвистам карты?

- показать, где работали
- показать распределение языкового использования в пространстве (и времени)
- показать распределение языковых признаков в пространстве (и времени)
- исследовать распределение сочетаемость языковых признаков в пространстве

Таким образом можно обнаружить

- что-то из истории изменения языков: контакты, завоевания и т. п.
- что-то из теоретических возможностей: какие языковые признаки могут сосуществовать внутри одной системы

ссылка на презентацию: tinyurl.com/ycx46od6

Glottolog

Адыгейский язык назван в литературе:

- *черкесскій* или *адигскій* — [Люлье 1846]
- *адыгейский язык* — [Рогава, Керашева 1966]
- *West Circassian* — [Smeets 1984]
- *Tcherkesse occidental* — [Paris, Batouka 2005]
- *Adyghe* — [Korotkova, Lander 2010]
- еще встречается *нижнечеркесский*, *западночеркесский*, *западноадыгский*, *кяхский*

ссылка на презентацию: tinyurl.com/ycx46od6

Glottolog

Адыгейский язык назван в литературе:

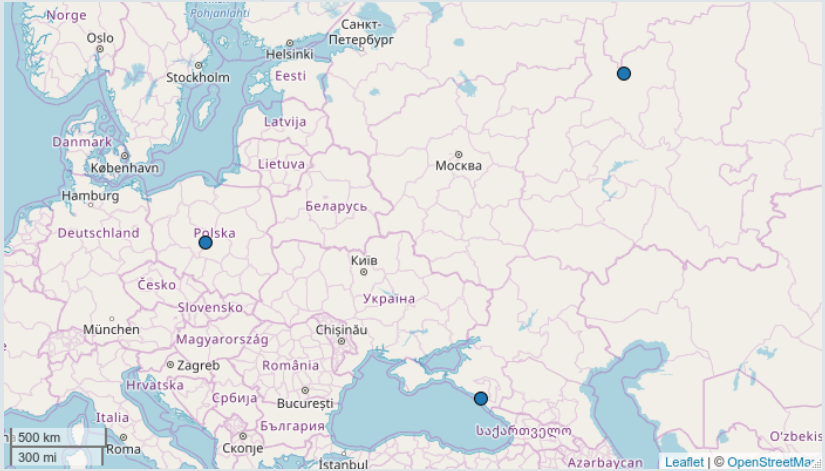
- *черкесскій* или *адигскій* — [Люлье 1846]
- *адыгейский язык* — [Рогава, Керашева 1966]
- *West Circassian* — [Smeets 1984]
- *Tcherkesse occidental* — [Paris, Batouka 2005]
- *Adyghe* — [Korotkova, Lander 2010]
- еще встречается *нижнечеркесский*, *западночеркесский*, *западноадыгский*, *кяхский*

На все это есть один код: `adyg1241`

ссылка на презентацию: tinyurl.com/ycx46od6

Glottolog + Leaflet

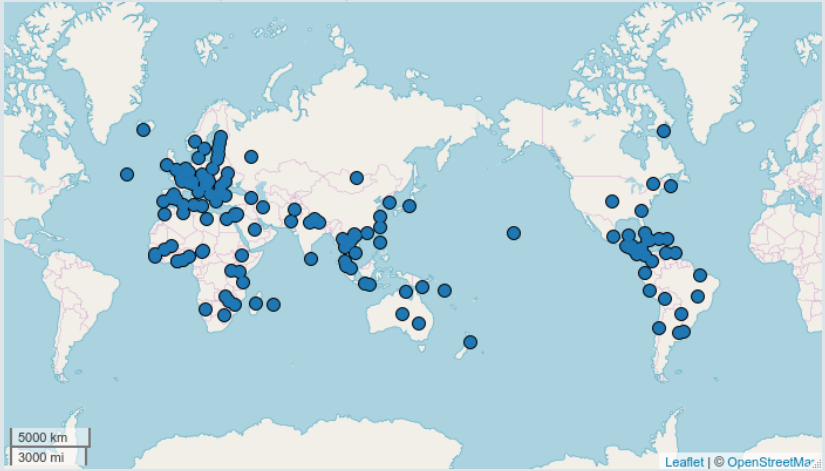
```
map.feature(c("Adyghe", "Russian", "Polish"))
```



ссылка на презентацию: tinyurl.com/ycx46od6

Glottolog + Leaflet

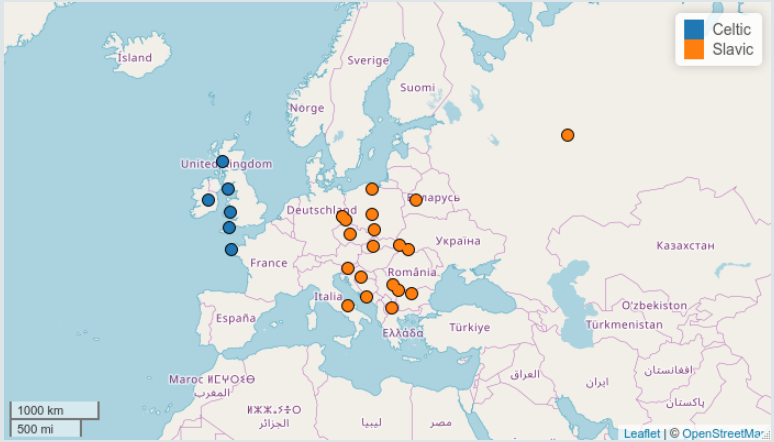
```
map.feature(lang.aff("Sign"))
```



ссылка на презентацию: tinyurl.com/ycx46od6

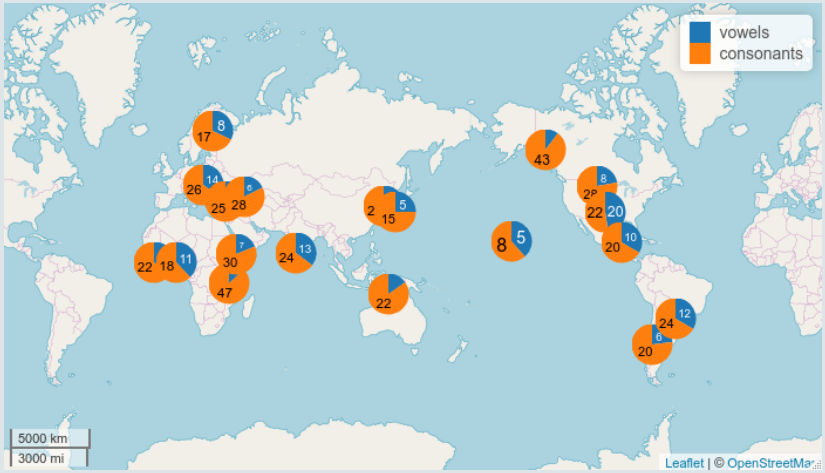
Glottolog + Leaflet

```
map.feature(c(lang.aff("Slavic"), lang.aff("Celtic")),
            c(rep("Slavic", 20), rep("Celtic", 6)))
```



ссылка на презентацию: tinyurl.com/ycx46od6

Glottolog + Leaflet

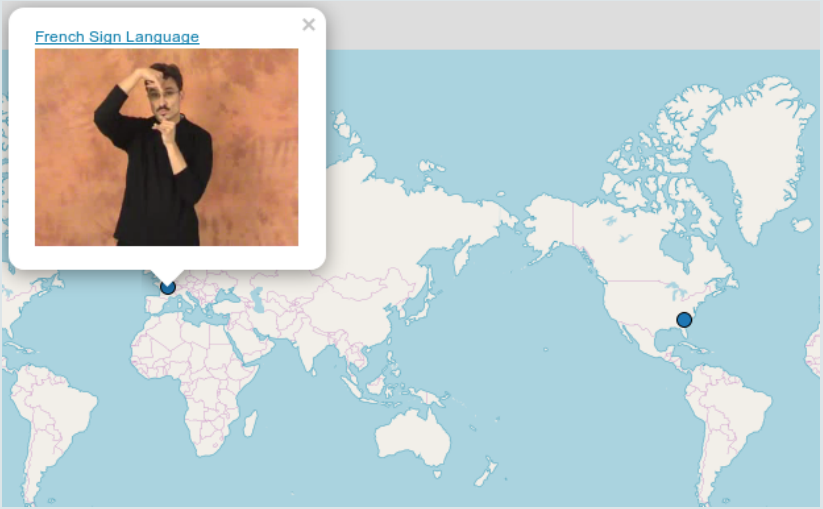


ссылка на презентацию: tinyurl.com/ycx46od6

Glottolog + Leaflet

ссылка на презентацию: tinyurl.com/ycx46od6

Glottolog + Leaflet



ссылка на презентацию: tinyurl.com/ycx46od6

Результаты:

- Более 14 тыс. скачиваний

ссылка на презентацию: tinyurl.com/ycx46od6

Результаты:

- Более 14 тыс. скачиваний
- В lingtypology внедрились API для других лингвистических баз данных

ссылка на презентацию: tinyurl.com/ycx46od6

Результаты:

- Более 14 тыс. скачиваний
- В lingtypology внедрились API для других лингвистических баз данных
- Множество редакций туториала; лекции и мастерклассы

ссылка на презентацию: tinyurl.com/ycx46od6

Результаты:

- Более 14 тыс. скачиваний
- В lingtypology внедрились API для других лингвистических баз данных
- Множество редакций туториала; лекции и мастерклассы
- Появляются базы данных, основанные на lingtypology:
  - The World Writing System Database (WWSD)
  - The World Consonant Alternation Database
  - Daghestanian Sound Database
  - The Circassian Isoglosses Database
  - Uvular consonants in Languages of the Caucasus
  - Iconicity patterns in Sign Languges
  - The Sound Change in Sibilants Database
  - Typological atlas of Guatemala

ссылка на презентацию: tinyurl.com/ycx46od6

Результаты:

- Более 14 тыс. скачиваний
- В lingtypology внедрились API для других лингвистических баз данных
- Множество редакций туториала; лекции и мастерклассы
- Появляются базы данных, основанные на lingtypology:
  - The World Writing System Database (WWSD)
  - The World Consonant Alternation Database
  - Daghestanian Sound Database
  - The Circassian Isoglosses Database
  - Uvular consonants in Languages of the Caucasus
  - Iconicity patterns in Sign Languges
  - The Sound Change in Sibilants Database
  - Typological atlas of Guatemala
- 6 issues на гитхабе Glottolog'а

ссылка на презентацию: tinyurl.com/ycx46od6

Результаты:

- Более 14 тыс. скачиваний
- В lingtypology внедрились API для других лингвистических баз данных
- Множество редакций туториала; лекции и мастерклассы
- Появляются базы данных, основанные на lingtypology:
  - The World Writing System Database (WWSD)
  - The World Consonant Alternation Database
  - Daghestanian Sound Database
  - The Circassian Isoglosses Database
  - Uvular consonants in Languages of the Caucasus
  - Iconicity patterns in Sign Languges
  - The Sound Change in Sibilants Database
  - Typological atlas of Guatemala
- 6 issues на гитхабе Glottolog'а
- рецензия rOpenSci

ссылка на презентацию: tinyurl.com/ycx46od6

Результаты:

- Более 14 тыс. скачиваний
- В lingtypology внедрились API для других лингвистических баз данных
- Множество редакций туториала; лекции и мастерклассы
- Появляются базы данных, основанные на lingtypology:
  - The World Writing System Database (WWSD)
  - The World Consonant Alternation Database
  - Daghestanian Sound Database
  - The Circassian Isoglosses Database
  - Uvular consonants in Languages of the Caucasus
  - Iconicity patterns in Sign Languges
  - The Sound Change in Sibilants Database
  - Typological atlas of Guatemala
- 6 issues на гитхабе Glottolog'a
- рецензия rOpenSci
- все больше появляется студенческих работ с картами

ссылка на презентацию: tinyurl.com/ycx46od6

Спасибо за внимание!

Пишите письма
agricolamz@gmail.com

Рисуйте карты с lingtypology