

Ольга
Ляшевская

Цифровые ресурсы для лингвистических исследований

Лекция 2



Курс
«Лингвистические данные»
НИУ ВШЭ, ФиКЛ, 1 курс бакалавриата

Сценарий 1



Сценарий 1

**На какие ресурсы следует опираться
при исследовании (неизвестного) языка?**

Сценарий 1



Документация языков:

**можно ли написать текст на незнакомом языке,
прочитав его грамматику и словарь?**

Сценарий 1 - документация языков

Можно ли написать текст на незнакомом языке,
прочитав его **грамматику и словарь?**

Нет - нужно прочитать много текстов, а еще
лучше, пообщаться с носителями языка.

- **язык - средство общения**
- **в языке всегда есть много вариантов выражения мысли - в зависимости от намерений говорящего и коммуникативной ситуации**
- **язык - живой, языковые средства могут меняться**

Откуда тексты?



Цифровая революция и лингвистика

Интернет как фонд текстов на языке N:

- новости, тексты из газет и журналов, электронные версии книг, сценарии кинофильмов, сайты музеев и учебных заведений, обзоры товаров, транскрипты интервью, реклама...
- аудио- и видеозаписи: радиопрограммы, интервью, аудиокниги, радиоспектакли, песни, youtube/gutube..
- социальные сети, форумы, чаты
- справочные, энциклопедические и образовательные ресурсы
- поисковые системы и переводчики

Откуда тексты?

Цифровая революция и лингвистика

книги

радио, телевидение

словари + энциклопедии

учебники

было

стало

электронные книги аудиокниги
интернет-тексты

электронные словари

электронные пособия,
медиакурсы, тренажеры
записи онлайн-обучения (Skype)
ресурсы для переводчиков

Откуда тексты?

Цифровая революция и лингвистика

книги

радио, телевидение

словари + энциклопедии

учебники

было

стало

электронные книги аудиокниги
интернет-тексты

электронные словари

электронные пособия,
медиакурсы, тренажеры
записи онлайн-обучения (Skype)
ресурсы для переводчиков

Сгенерированные тексты?

Электронные библиотеки (примеры)

- Google Books - books.google.com
- Universal Digital [Library](#)
- Проект Гутенберг - www.gutenberg.org
- Internet [Archive.org](#)
- "Народные" проекты
 - [lib.ru](#) Библиотека Максима Мошкова
 - [netslova.ru](#) Сетевая словесность
 - [russ.ru](#) Русский журнал, [stihi.ru](#) Стихи.ру...
- Академические проекты
- [feb-web.ru](#) Фундаментальная электронная библиотека "Русская литература и фольклор" – аннотированные электронные версии классики, включая варианты изданий (там же словари и литературные энциклопедии)
- [wikipedia.org](#) Википедия (архив как большой текстовый ресурс)



Сценарий 1 - документация языков

Корпус - лучше, чем текстовый архив

Корпус – коллекция текстов, снабженная специальной разметкой (информация о самих текстах, о каждом предложении и слове)

Электронные корпуса



Типичные вопросы, на которые отвечают корпуса:

- отличается ли речь авторов-женщин от авторов-мужчин?
- когда впервые появилось в языке слово слямзить?
(NB! не появилось, а задокументировано)
- отличается ли сочетаемость слов хотеть и стремиться?
(ср. *?я стремился, чтобы...*)

Электронные корпуса



Типичные вопросы, на которые отвечают корпуса:

- отличается ли речь авторов-женщин от авторов-мужчин?
 - <все тексты должны иметь помету "пол автора">
- когда впервые появилось в языке слово слямзить?
(NB! не появилось, а задокументировано)
 - <все тексты должны иметь помету "дата создания">
 - <корпус должен уметь находить слово во всех формах - разметка лексем>
- отличается ли сочетаемость слов хотеть и стремиться?
(ср. ?я стремился, чтобы...)
 - <синтаксическая разметка - связь и расстояние между словами>

Электронные корпуса



Собираются на основе существующих текстов, но включают добавленное знание (added knowledge)

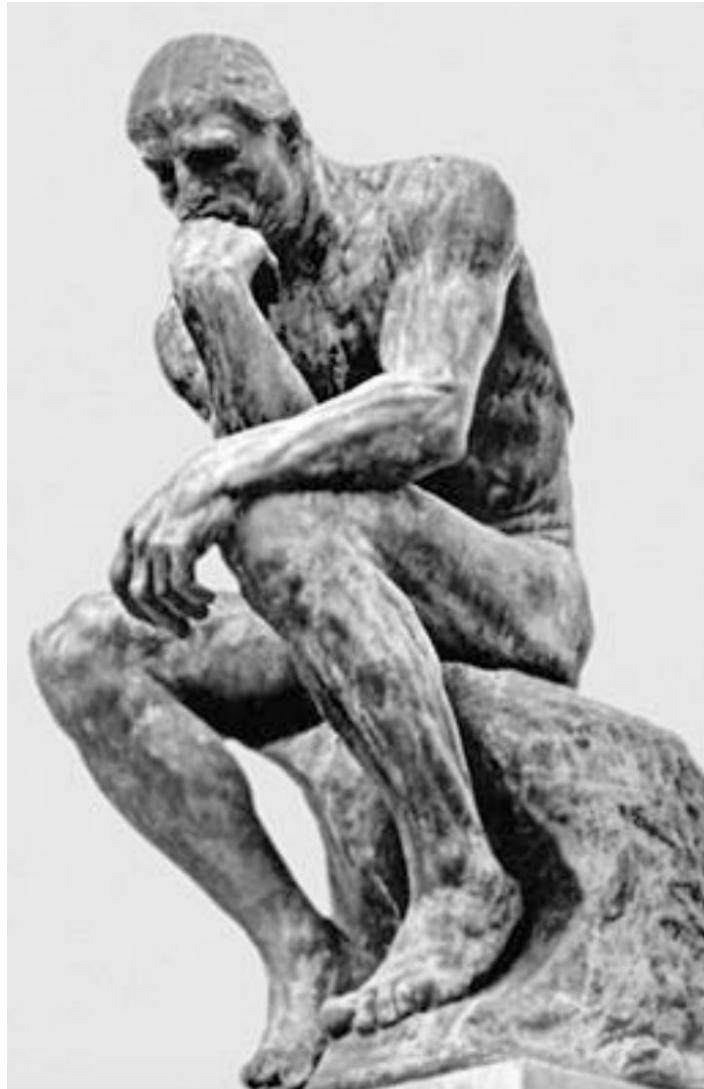
- лингвистическое
 - про каждый звук, букву/иероглиф, слог, слово, предложение, абзац и так далее - в зависимости от той или иной теории
- энциклопедическое
 - разметка топонимов, годы жизни авторов и т.п.
- стиховедческое / литературоведческое
 - метрика, виды рифмы, типы элегий и т.п.
- историческое, Computer Science/NLP/AI, психологическое...

Классификация ресурсов

- грамматики
 - корпуса
 - словари
 - справочные системы ([Грамота.ру](#))
 - другие специальные ресурсы (GIS-системы для диалектных исследований, [common voice](#) и т.п.)
- базы данных
- грамматические БД - структурированные факты по грамматикам
 - надкорпусные БД,
в т.ч. частотные
 - лексикографические БД - структурированные факты о лексике



Что еще?



Чем еще
пользуются
лингвисты?



Что еще?

- языковая интуиция?
 - если исследователь сам носитель языка, интуиция поможет ему ответить на вопрос, существует ли или иная конструкция/единица
-
- если нет... можем спросить **других носителей!**
- собрать непрямые свидетельства
- и какого типа у нас получатся ресурсы?



Ресурсы для/как результат социо-, психо-, нейролингвистических, полевых и т.п. исследований

- унифицированные анкеты/опросники
- сбалансированные базы стимулов
- наборы данных (ответы, измерения времени
реакции, ЭЭГ...)
корпусные наборы данных (выборки)
- библиографии / mind maps
 - воспроизводимость исследований



Примеры



NATIONAL RESEARCH
UNIVERSITY

Корпуса

- Национальные корпуса - BNC, НКРЯ...
- Интернет-корпуса - EnTenTen, RuTenTen, Wacky, Aranea...
- Мониторинговые - газетный корпус НКРЯ
- Диахронические - СОНА
- Устные корпуса
- Мультимедийные - The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s),
- Параллельные - OPUS project, параллельные корпуса НКРЯ
- Сопоставимые корпуса - CHILDES TalkBank, Wikipedia как корпус
- Дialectные корпуса - HSE LingConLab согрода
- Учебные & эритажные корпуса - LINDSEI (Louvain International Database of Spoken English Interlanguage), REALEC



NATIONAL RESEARCH
UNIVERSITY



ekje ekk ekke ekkj ekke ennte ente ett ette gge iije ikj ikje ikk ikka ikke ikket ikki ikkj ikkje
ikkjee ikkji ikkke ikkkje ikø inggjke ingkji inngkje inngkji innkje innkji innt innte inntje inte issj issje
issjæ it itj itje itt itte itti ittj ittjæ ittse itsje je ke kj kje kjj kjæ kk kke kkj kkje kkji
nnte nte rrt rrte rt rte si sj sje sjæ ssje t te tj tje tkje tne tsje tt tte ttje ænnte ætte



Nordic Dialect Corpus:
Search result for the word
ikke ‘not’ (a map view).

Базы данных

- Типологические

7,000 living languages:
families, location & maps,
population size, dialects

Ethnologue

WALS

World Atlas of Language
Structures

languages & languoids,
incl. sign languages and
artificial languages

Glottolog

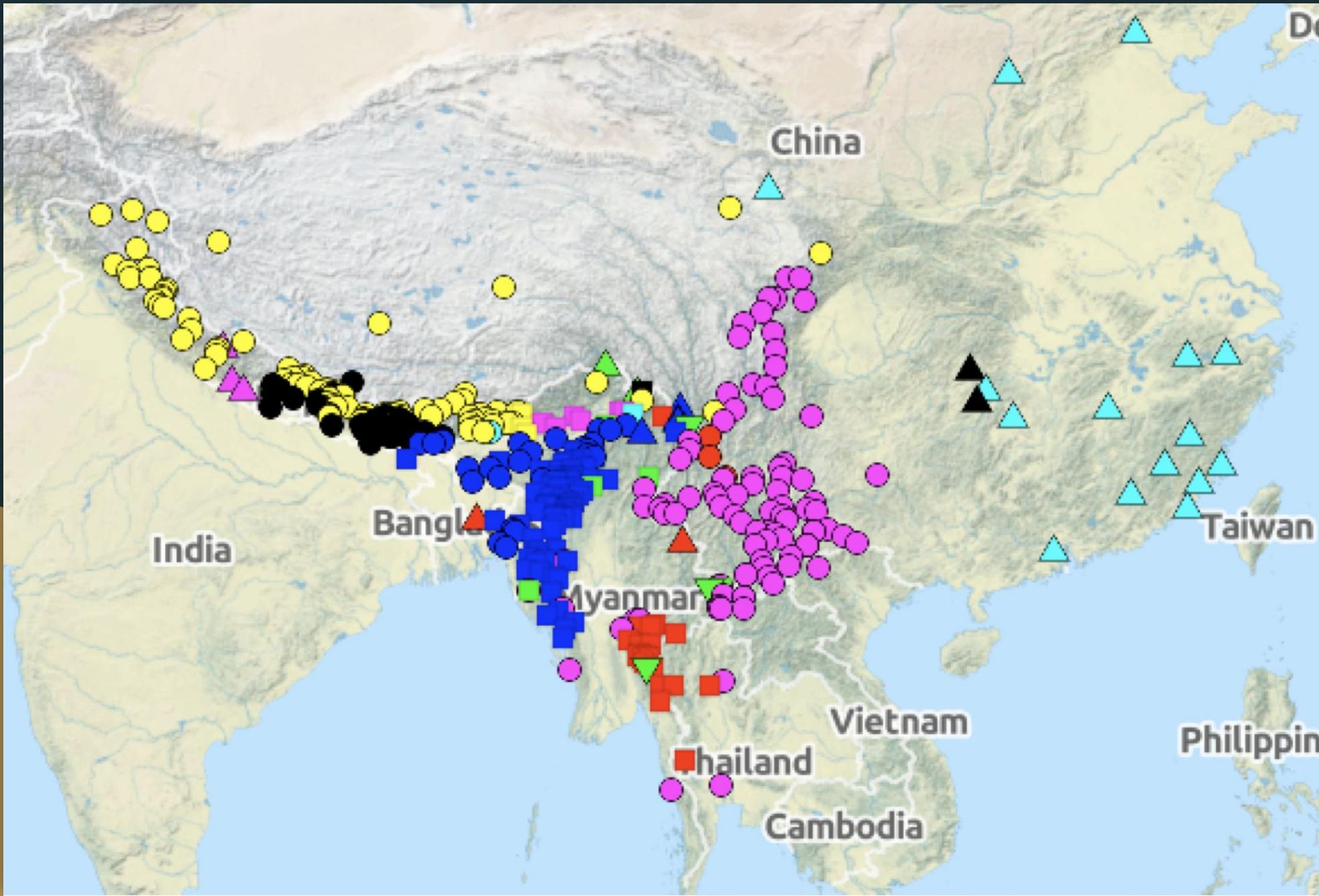
PHOIBLE

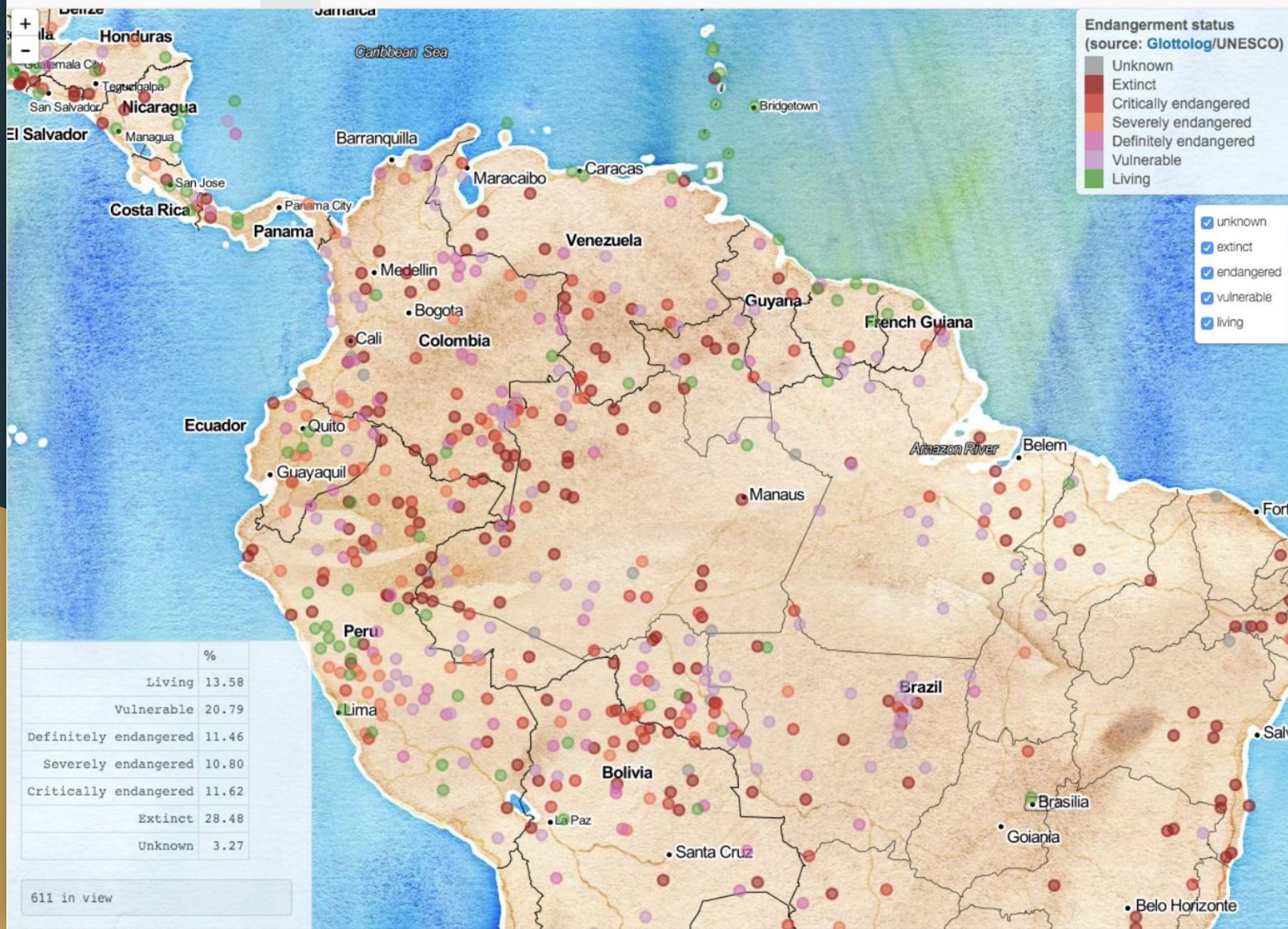
phonological
inventories



NATIONAL RESEARCH
UNIVERSITY

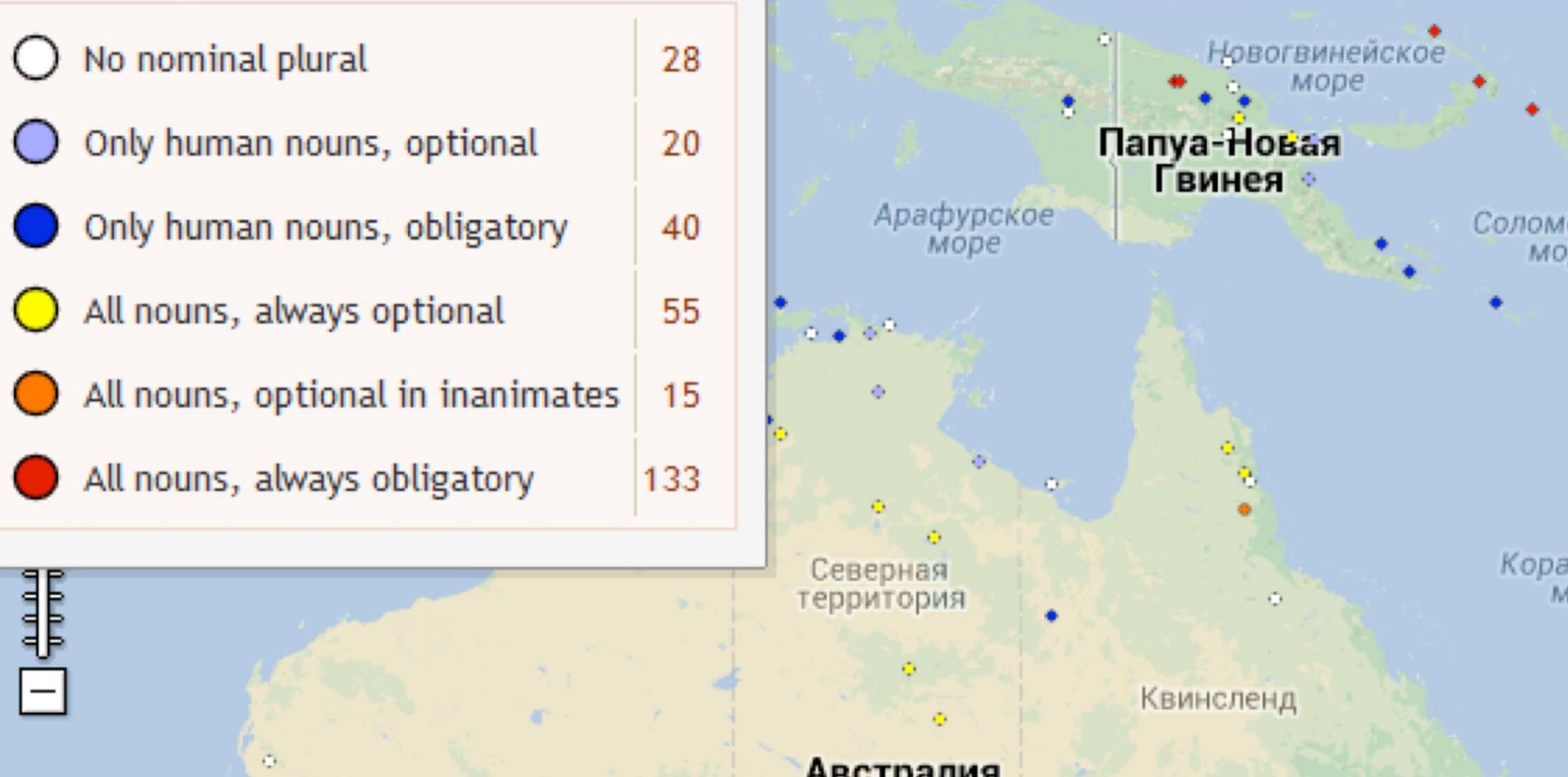
Glottolog:
Geographical
distribution of
Sino-Tibetan
languages





Glottolog data explorer map:
The northern region of South America including the Amazon rainforest
([Caines et al. 2016](#))

- | | |
|-------------------------------------|-----|
| ○ No nominal plural | 28 |
| ● Only human nouns, optional | 20 |
| ● Only human nouns, obligatory | 40 |
| ● All nouns, always optional | 55 |
| ● All nouns, optional in inanimates | 15 |
| ● All nouns, always obligatory | 133 |

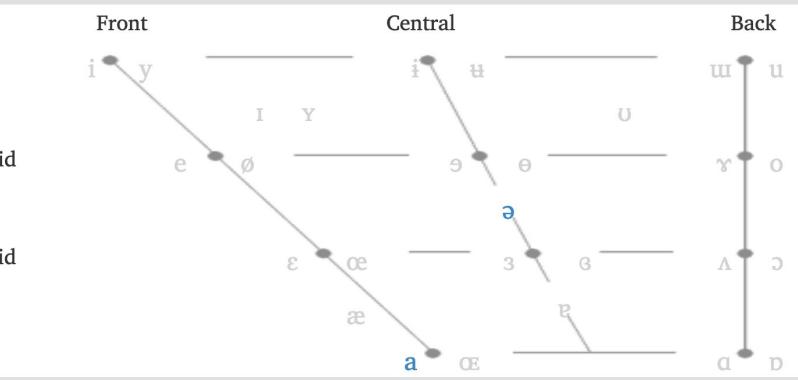


WALS: Feature 34A: Occurrence of Nominal Plurality

Consonants (Pulmonic)

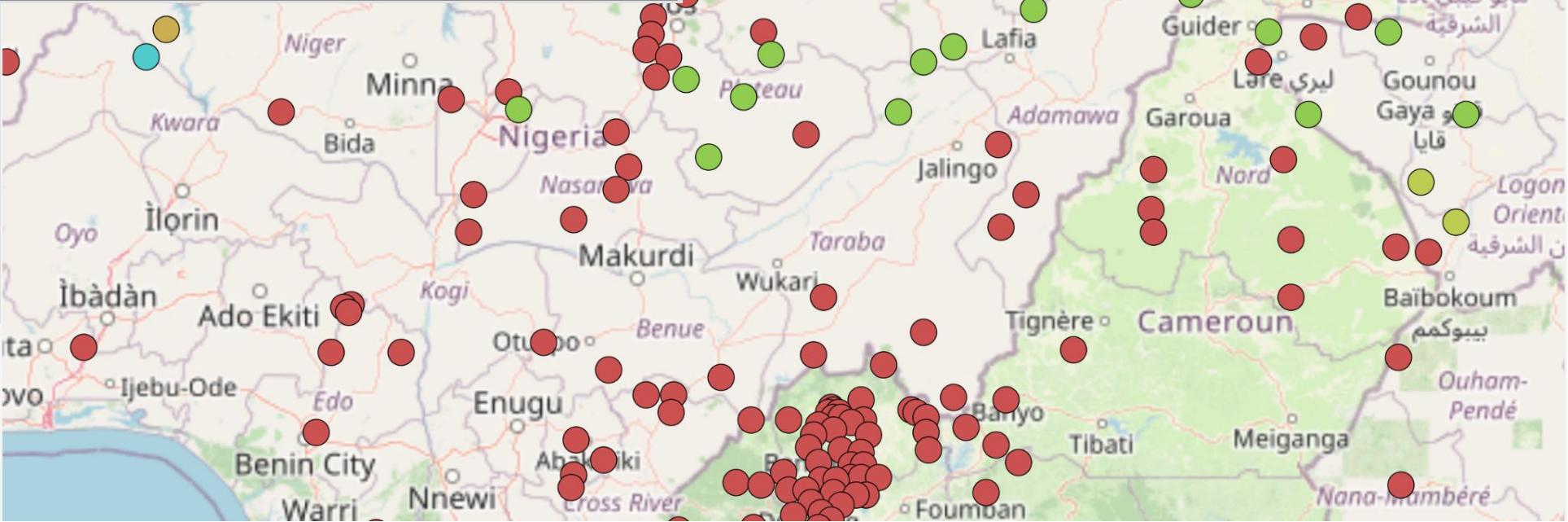
| | Bilabial | Labiodental | Dental | Alveolar | Postalveolar | Retroflex | Palatal | Velar | Uvular | Pharyngeal | Glottal |
|---------------------|----------|-------------|--------|----------|--------------|-----------|---------|-------|--------|------------|---------|
| Plosive | p b | | t d | t d | | t d | c j | k g | q G | | ? |
| Nasal | m | nj | n̪ | n | | n̪ | n̪ | ŋ | N | | |
| Trill | r | | r̪ | | | | | R | | | |
| Tap or Flap | v | f̪ | r̪ | r̪ | t̪ | t̪ | | | | | |
| Fricative | ɸ β | f v | θ ð | s z | ʃ ʒ | ʂ ʐ | ç ɿ | x ɣ | χ ʁ | h χ | h f̪ |
| Lateral fricative | | | ɬ ɺ | | | | | | | | |
| Approximant | | v̪ | | i̪ | | l̪ | j̪ | w̪ | | | |
| Lateral approximant | | | l̪ | l̪ | | l̪ | ɻ̪ | ɻ̪ | | | |

Vowels



Consonants (Non-Pulmonic)

| Clicks | Voiced implosives |
|------------------|-------------------|
| ○ Bilabial | ɓ Bilabial |
| — Dental | ɗ Dental/alveolar |
| (Post)alveolar | ʄ Palatal |
| + Palatoalveolar | ɠ Velar |
| Alveolar lateral | ʄ Uvular |

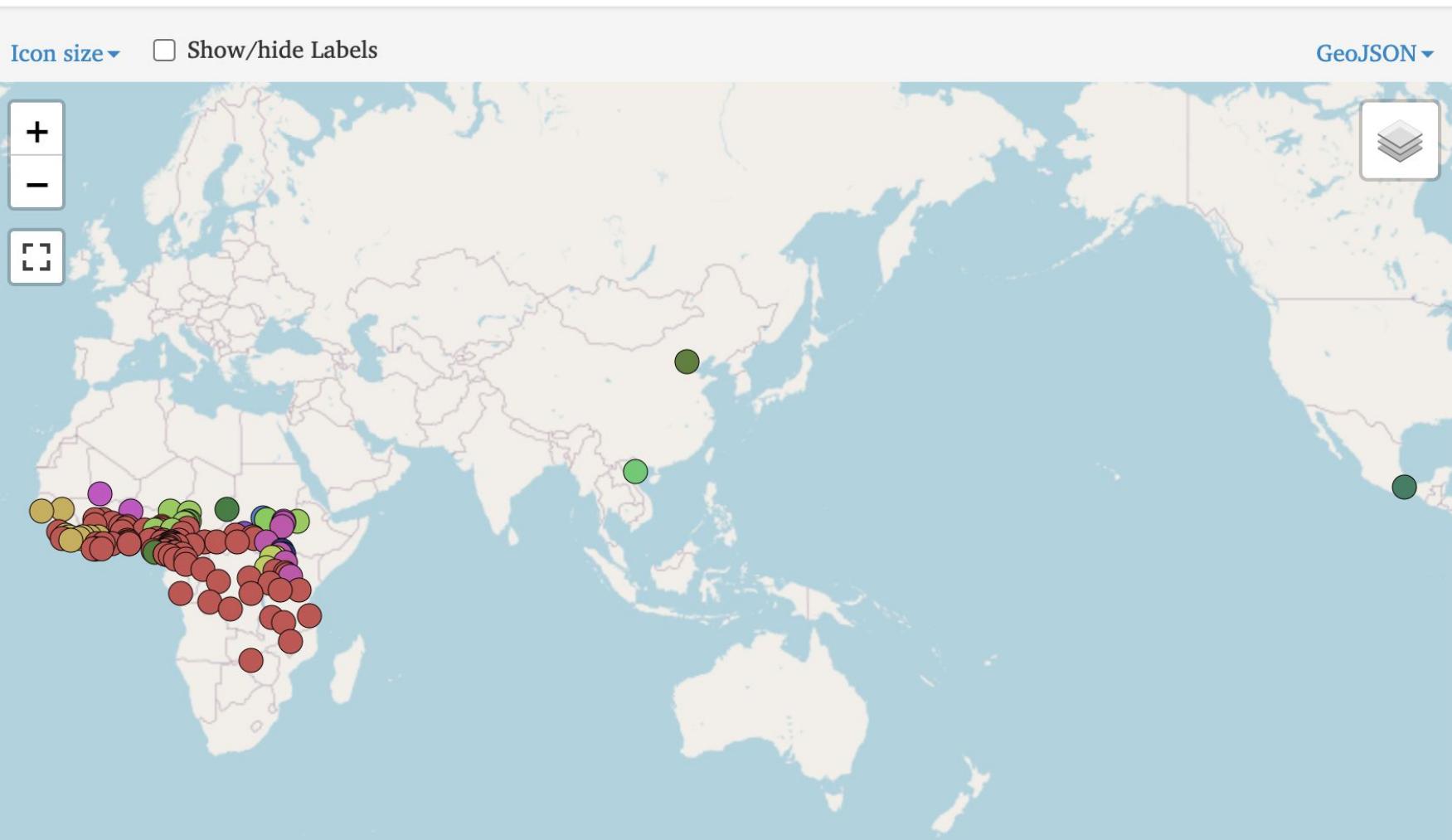


PHOIBLE: **zulgo** inventory: 2 vowels, 34 consonants

Tone \ W



MODIFIER LETTER HIGH TONE BAR - MODIFIER LETTER LOW TONE BAR



PHOIBLE: Languages with tone \ - MODIFIER LETTER HIGH TONE BAR - MODIFIER LETTER LOW TONE BAR

Базы данных

- **Лексические**

etymological database

The Tower
of Babel

117,000 synsets
(synonyms, hyperonyms
& hyponyms, meronyms,
antonyms, etc.)

VisuWords

lexicon as a net:
visual dictionary,
visual thesaurus,
interactive lexicon

WordNet

Common
Voice

pronunciation
crowdsourcing
project



NATIONAL RESEARCH
UNIVERSITY

Number: 1713

Proto: *wasa

English meaning: calf, deer calf

German meaning: Kalb, Renkalb

Finnish: vasa 'Kalb, einjähriges Renkalb', vasikka 'Kalb'

Estonian: vasik, vasikas (gen. vasika)

Saam (Lapp): vyesi (I), vi^šisse (T), vū^šss (Kld.), vuai^šss (Not.) 'kleines Rentierkalb, bis es um den Peterstag neues Haar bekommt'

Mordovian: vaz (E M), vazńe (E), vazńä (M) 'Kalb'

Mansi (Vogul): (wēsəj KM, wēsəy P, wāsiy So. 'Elchkalb' - rejected by Redei as "eine vom FW unabhängige iran. Entlehnung")

Гибридные ресурсы

frames as predicate & arguments structures: semantic and syntactic roles, adjuncts, frame scenarios

- **Syntax, semantics, collocations**

automatic corpus-based summary of a word's grammatical and collocational behaviour

FrameNet

Constructicon

Sketch Engine

Trados

multiword constructions (core grammatical structures & idiomatic, non-transparent form-meaning pairings)

computer-assisted translation (CAT) tool: translation memory & term database



NATIONAL RESEARCH
UNIVERSITY

goal

(noun) ukWaC freq = 168345 (107.5 per million)

| <u>object of</u> | 58924 | 3.2 | <u>subject of</u> | 25451 | 2.4 | <u>modifier</u> | 67879 | 1.6 | <u>modifies</u> | 11026 | 0.3 |
|------------------|-----------------------|------------|-------------------|-----------------------|------------|-----------------|-----------------------|------------|-----------------|-----------------------|------------|
| score | 8390 | 11.28 | score | 903 | 8.59 | ultimate | 1911 | 9.27 | scorer | 389 | 9.39 |
| achieve | 9422 | 9.9 | disallow | 223 | 8.04 | long-term | 875 | 7.66 | kick | 634 | 8.86 |
| concede | 1421 | 9.39 | concede | 204 | 7.53 | league | 638 | 7.38 | tally | 129 | 7.9 |
| accomplish | 585 | 7.97 | gape | 76 | 6.5 | winning | 401 | 7.33 | keeper | 204 | 7.31 |
| reach | 1924 | 7.66 | come | 1316 | 5.44 | primary | 993 | 7.24 | scramble | 50 | 6.75 |
| net | 337 | 7.42 | kick | 76 | 5.44 | second | 2000 | 7.19 | drought | 78 | 6.65 |
| pursue | 648 | 7.41 | rule | 61 | 5.24 | common | 1529 | 7.17 | difference | 676 | 6.28 |
| attain | 400 | 7.35 | orientate | 34 | 5.06 | strategic | 645 | 7.1 | cushion | 53 | 6.26 |
| grab | 406 | 7.34 | arrive | 90 | 4.43 | realistic | 422 | 7.05 | lead | 267 | 6.24 |
| set | 2413 | 7.01 | cap | 20 | 4.38 | achievable | 290 | 6.97 | setting | 405 | 6.14 |
| pull | 501 | 6.88 | beat | 53 | 4.31 | stated | 259 | 6.8 | kicker | 25 | 6.04 |
| disallow | 190 | 6.67 | direct | 53 | 4.22 | score | 611 | 6.75 | post | 482 | 5.91 |

clever/intelligent

ukWaC freqs = 20589/26115

| | | | | | | | | |
|--------|-----|-----|-----|---|------|------|------|-------------|
| clever | 6.0 | 4.0 | 2.0 | 0 | -2.0 | -4.0 | -6.0 | intelligent |
|--------|-----|-----|-----|---|------|------|------|-------------|

| and/or | 4955 | 10062 | 2.2 | 3.6 | modifier | 4950 | 3168 | 0.9 | 0.5 | modifies | 10948 | 16081 | 2.0 | 2.4 |
|-------------------|------|-------|-----|-----|-----------------|------|------|-----|-----|--------------|-------|-------|-----|-----|
| perceptive | 0 | 34 | 0.0 | 6.4 | emotionally | 0 | 111 | 0.0 | 8.6 | being | 0 | 208 | 0.0 | 6.1 |
| thought-provoking | 0 | 32 | 0.0 | 6.2 | artificially | 0 | 52 | 0.0 | 7.9 | robot | 0 | 77 | 0.0 | 6.1 |
| informed | 0 | 66 | 0.0 | 6.2 | fiercely | 0 | 26 | 0.0 | 7.0 | agent | 9 | 455 | 0.4 | 6.0 |
| autonomous | 0 | 46 | 0.0 | 6.2 | highly | 0 | 570 | 0.0 | 6.9 | guess | 0 | 35 | 0.0 | 5.5 |
| adaptive | 0 | 39 | 0.0 | 6.1 | ferociously | 0 | 8 | 0.0 | 6.2 | routing | 0 | 27 | 0.0 | 5.3 |
| well-informed | 0 | 24 | 0.0 | 6.0 | supposedly | 0 | 28 | 0.0 | 6.2 | layman | 0 | 22 | 0.0 | 5.3 |
| literate | 0 | 26 | 0.0 | 5.9 | averagely | 0 | 7 | 0.0 | 6.1 | conversation | 0 | 88 | 0.0 | 5.1 |
| compassionate | 0 | 27 | 0.0 | 5.9 | moderately | 0 | 11 | 0.0 | 5.7 | creature | 11 | 137 | 2.4 | 5.9 |
| well-educated | 0 | 17 | 0.0 | 5.7 | reasonably | 0 | 54 | 0.0 | 5.7 | lyric | 81 | 80 | 5.8 | 5.7 |
| cultured | 0 | 19 | 0.0 | 5.7 | computationally | 0 | 6 | 0.0 | 5.6 | fellow | 52 | 14 | 5.1 | 3.1 |
| rational | 0 | 46 | 0.0 | 5.6 | supremely | 0 | 7 | 0.0 | 5.5 | pass | 67 | 9 | 5.2 | 2.2 |
| playful | 0 | 22 | 0.0 | 5.6 | culturally | 0 | 12 | 0.0 | 5.5 | stuff | 146 | 6 | 5.1 | 0.4 |
| sensitive | 8 | 134 | 2.0 | 5.9 | exceptionally | 29 | 25 | 6.0 | 5.9 | gimmick | 15 | 0 | 5.1 | 0.0 |
| thoughtful | 14 | 121 | 5.0 | 7.7 | remarkably | 24 | 11 | 5.7 | 4.8 | satire | 19 | 0 | 5.1 | 0.0 |
| affectionate | 6 | 31 | 4.5 | 6.2 | amazingly | 17 | 7 | 5.9 | 5.0 | flick | 21 | 0 | 5.2 | 0.0 |
| sophisticated | 23 | 75 | 4.2 | 5.7 | wonderfully | 20 | 9 | 5.4 | 4.5 | lob | 15 | 0 | 5.3 | 0.0 |
| charming | 21 | 50 | 4.9 | 5.9 | very | 1707 | 596 | 5.6 | 4.0 | pun | 19 | 0 | 5.3 | 0.0 |
| insightful | 11 | 31 | 5.2 | 6.1 | too | 476 | 76 | 5.4 | 2.8 | ruse | 17 | 0 | 5.5 | 0.0 |
| resourceful | 12 | 29 | 5.8 | 6.3 | damn | 12 | 0 | 5.6 | 0.0 | eh | 24 | 0 | 5.8 | 0.0 |
| witty | 132 | 166 | 8.3 | 8.2 | dead | 16 | 0 | 5.8 | 0.0 | wordplay | 21 | 0 | 5.8 | 0.0 |
| inventive | 22 | 24 | 5.9 | 5.5 | diabolically | 9 | 0 | 5.9 | 0.0 | chap | 47 | 0 | 5.9 | 0.0 |
| clever | 54 | 30 | 5.8 | 4.8 | awfully | 15 | 0 | 6.1 | 0.0 | twist | 94 | 0 | 6.5 | 0.0 |
| funny | 233 | 103 | 7.0 | 5.7 | terribly | 25 | 0 | 6.2 | 0.0 | trick | 166 | 0 | 6.7 | 0.0 |
| cunning | 16 | 7 | 5.9 | 4.0 | devilishly | 17 | 0 | 6.8 | 0.0 | clog | 50 | 0 | 7.0 | 0.0 |
| catchy | 19 | 0 | 5.8 | 0.0 | fiendishly | 45 | 0 | 8.1 | 0.0 | ploy | 68 | 0 | 7.2 | 0.0 |

SketchEngine:
contrastive sketch
profiles for the
adjectives *clever*
and *intelligent*

СЛОВНИКИ

Swadesh lists for
historical studies of
languages

Concepticon

Questionnaires

for linguistic field work
(e.g. emotions, body
parts, spacial relations)

frequency-balanced word
lists for experimental
studies

Naming
tests



NATIONAL RESEARCH
UNIVERSITY

А есть ли у вас идеи?



NATIONAL RESEARCH
UNIVERSITY

Заключение

- “The contents of a corpus [and other resources!] should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.”
John Sinclair 2004
- Конвертация (традиционных) словарей и грамматик в формат структурированных баз данных, статистика и визуализации позволяют исследователям, учащимся и др. быстрее знакомиться с информацией и релевантными исследованиями
- Все ресурсы должны строиться согласно тщательно выработанному дизайну (выборки, баланс и т.п.) и подробно документироваться!
- Cross-Linguistic Linked (Open) Data movement облегчают доступ к ресурсам, распределенным по всему миру, и их интеграцию



NATIONAL RESEARCH
UNIVERSITY