

Ольга
Ляшевская

Цифровые ресурсы для лингвистических исследований

Лекция 2



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

На первой лекции...

Форматы документации (лингвистических) данных

Кодировки текстов



На первой лекции...

Форматы документации (лингвистических) данных
Unicode (UTF-8, UTF-16), ASCII, Windows-1251

Кодировки текстов
HTML, YAML, XML, JSON...



На первой лекции...

Форматы документации

Unicode (UTF-8, UTF-16), ASCII, Windows-1251

Кодировки текстов

HTML, YAML, XML, JSON...

	1DF0	1DF1	1DF2	...	1DFFF
0	ᆿ	ᆺ	ᆻ		
1	ᆼ	ᆽ	ᆾ		
2	ᆿ	ᆺ	ᆻ		
3	ᆼ	ᆽ	ᆾ		
4	ᆿ	ᆺ	ᆻ		
5	ᆼ	ᆽ	ᆾ		
6	ᆿ	ᆺ	ᆻ		
7	ᆼ	ᆽ	ᆾ		
8	ᆿ	ᆺ	ᆻ		
9	ᆼ	ᆽ	ᆾ		
A	ᆿ	ᆺ	ᆻ		
B	ᆼ	ᆽ	ᆾ		
C	ᆿ	ᆺ	ᆻ		
D	ᆼ	ᆽ	ᆾ		
E	ᆿ	ᆺ	ᆻ		
F	ᆼ	ᆽ	ᆾ		

На первой лекции...

Сериализация

- представление информации в виде более низкоуровневой структуры
 - формат хранения
 - битовая последовательность



Инструменты для разметки

Разметка (аннотация) – добавление информации к первоисточнику

делают эксперты (в том числе студенты и волонтеры) и машины (в том числе нейросетевые модели и скрипты, основанные на правилах)

связана с интерпретацией (анализом) данных



Инструменты для разметки: примеры

Разметка текстов

[Praat](#) – разметка звука

[ELAN](#) – для аудио и видео

[Inception](#), [GATE](#) – для текстов

(адаптированы для компьютерно-лингвистических задач)

[Fieldworks](#) – для полевых лингвистов



Примеры: Praat

Command
menu



Waveform



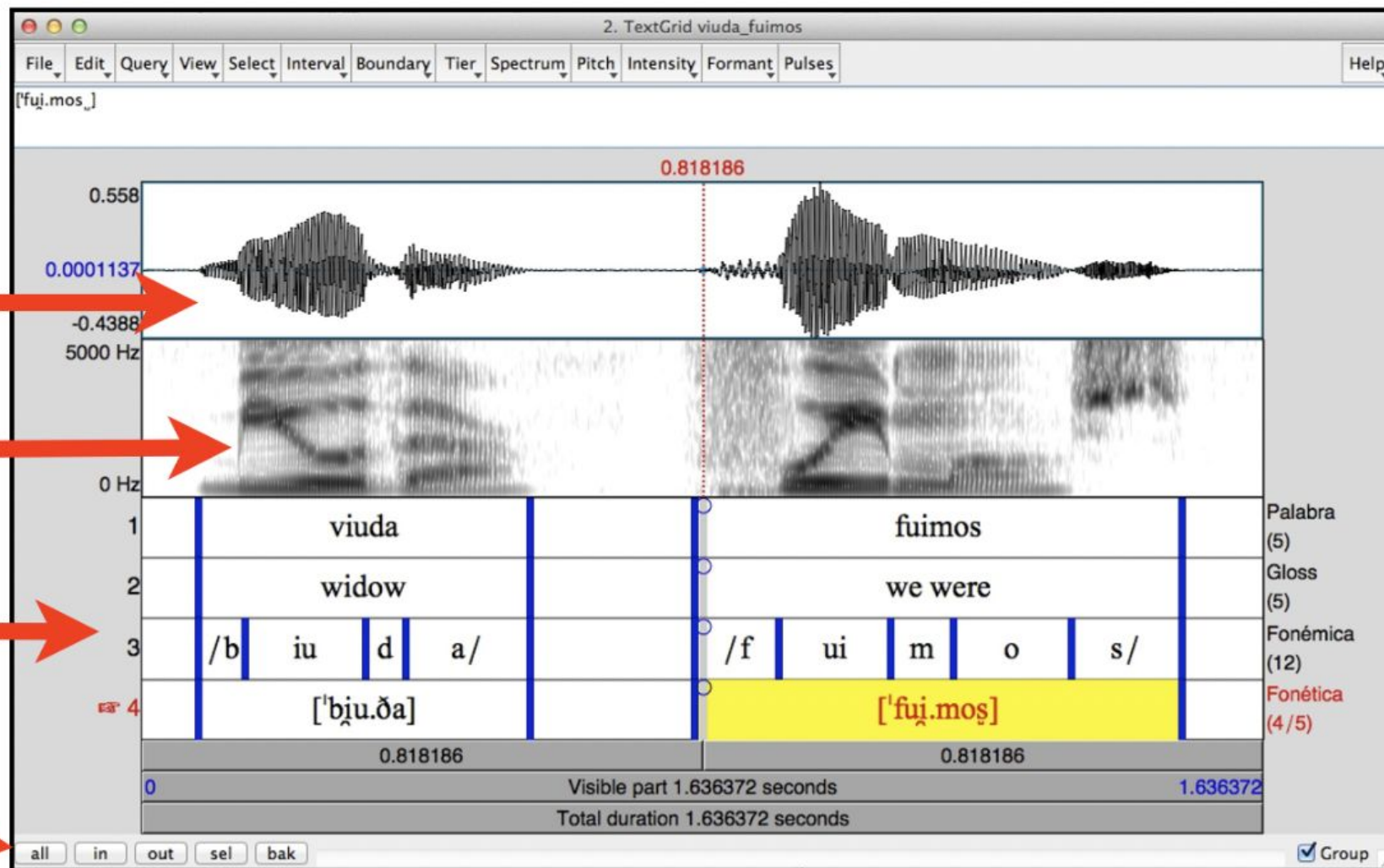
Spectrogram



Textgrid
analysis



Selection menu



Примеры: ELAN

ELAN - merged-utt.eaf

File Edit Annotation Tier Type Search View Options Window



Volume: 100

Rate: 100

00:06:26.950 Selection: 00:01:07.590 - 00:01:07.600 10

Selection Mode Loop Mode

00:06:25.000 00:06:30.000 00:06:35.000 00:06:40.000

RobotUtt [52] an tell me So, I am Arne (200) Which painting ha Great, your a Christian

Person1Utt [142]

Person2Utt [93] "speech" "silence" "speech" "silence"

Примеры: ELAN

ELAN 3.9.1 File Edit Annotation Tier Type Search View Options Window Help

Elan - pear story.eaf

Grid Text Subtitles Audio Recognizer Video Recognizer Metadata Controls

Recognizer: Tag vowels (volume peaks of voiced timespans) File(s): ☒ pear.wav

Selections

Add Remove Add Tier...

Parameters

Pitch ceiling [Hz], eg. 300 male / 550 female (150.0 - 1500.0) 604.8

Intensity change [dB] to start/end a peak (0.5 - 5.0) 2.0

Minimum amplitude (0..1) for pitch analysis (0.01 - 0.3)

Segmentations

Create Tier(s)...

Progress

Ready Start Report...

00:00:16.840 Selection: 00:00:00.000 - 00:00:00.000 0

78.005 55.3015 32.598 604.475 340.3425 76.21

38.9515 504.576

00 00:00:09.000 00:00:10.000 00:00:11.000 00:00:12.000 00:00:13.000 00:00:14.000 00:00:15.000 00:00:16.000 00:00:17.000 00:00:18.000 00:00:19.000 00

ent

ause Transcri

and so he climbs up a tree and he starts with the ladder and he starts picking pears off the tree and he puts the pears into an apron

motion non-motion non-motion

gesture #

gesture 1 gesture 2 gesture 3 gesture 4 gesture 5 gesture 6 gesture 7 gesture 8 gesture 9 gesture 10

3s Hand

R R R R R R R B B B

Инструменты для разметки: примеры

Разметка текстов

[Praat](#) – разметка звука

[ELAN](#) – для аудио и видео

[Inception](#), [GATE](#) – для текстов

[Fieldworks](#) – для полевых

Разработка словарей

iLex – разработка

[TLex](#) – разработка с AI-поддержкой

[Lexique PRO](#) – для публикации
Tickbox Lexicography (TBL)



Примеры: TLex для словарей

TLex - [C:\TshwaneDJe\Projects\PanSALB\Sesotho sa Leboa]

File Edit View Lemma Dictionary Format Tools Window

+ New entry (Ins) Delete entry

setlaela

setla [2]
setlabakelo
setlabelo
setlaela
setlamara
setlamo
setlanyi
setlatla
setlautlau
setleetlee
setlepetlepe
setlono

setlaela
setla [1]
setaele
se

Lemma: setlaela LemmaSign=setlaela,Pro
References: See lešilo

Attributes (F1) Attributes (F2)

Lemma: ☐ Incomplete

LemmaSign	setlaela
Pronunciation	setlaêla
Deriv	
Etymology	
Notes	
Frequency	0
References	

B I U S 2 2 Search

Sesotho sa Leboa [A B C D E F G H I J K L M N O P Q R S T U V W X Y Z]

setlaela /setlaêla/ (noun) (7/8) BONA lešilo

Article has cross-references to >>

lešilo /lešilô/ (noun) (5/6) MAHLALOŠETŠAGOTEE
setlaela, seota, senyabanyaba, setlamara,
setlatla, setlautlau
motho yo a sego a hlalefa, yoo gantši a dirago dilo tša
bobjana le go se kgone go fahlogela dilo tšeo di lego
molaleng : *Ke be ke ikwa bjalo ka lešilo la mafelelo
leo le sa tsebego selo le bilego le lebetše le ka
leina la lona*

>> Article is cross-referenced from

ditlaela /ditlaêla/ (pl noun) (7/8) BONA setlaela
lešilo /lešilô/ (noun) (5/6) MAHLALOŠETŠAGOTEE
setlaela, seota, senyabanyaba, setlamara,
setlatla, setlautlau
motho yo a sego a hlalefa, yoo gantši a dirago dilo tša
bobjana le go se kgone go fahlogela dilo tšeo di lego
molaleng : *Ke be ke ikwa bjalo ka lešilo la mafelelo
leo le sa tsebego selo le bilego le lebetše le ka
leina la lona*

Инструменты для разметки: примеры

Разметка текстов

[Praat](#) – разметка звука

[ELAN](#) – для аудио и видео

[Inception](#), [GATE](#) – для текстов

[Fieldworks](#) – для полевых

Разработка корпусов

KonText, Corpus Workbench
конкордансеры

+ текстовые редакторы, редакторы баз данных...

Разработка словарей

iLex – разработка

[TLex](#) – разработка с AI-поддержкой

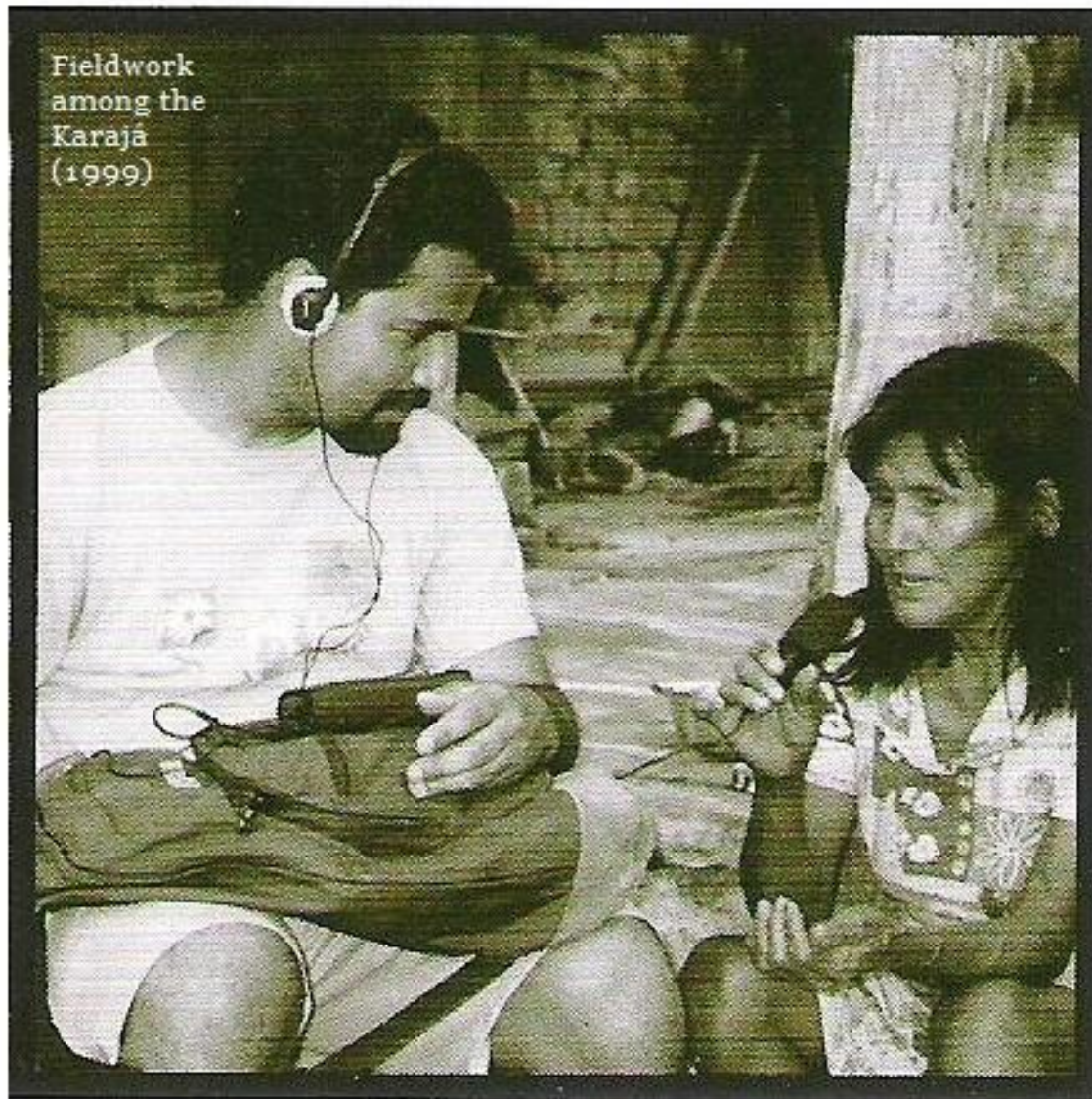
[Lexique PRO](#) – для публикации
Tickbox Lexicography (TBL)

Перевод

Translation memory tools



Документация языков



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*

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

стало

книги

электронные книги аудиокниги

радио, телевидение

интернет-тексты

словари + энциклопедии

электронные словари

учебники

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

было

ресурсы для переводчиков

Откуда тексты?

Цифровая революция и лингвистика

стало

книги

электронные книги аудиокниги

радио, телевидение

интернет-тексты

словари + энциклопедии

электронные словари

учебники

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

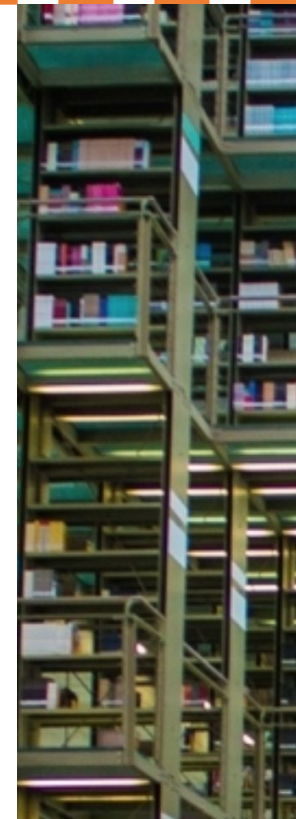
было

ресурсы для переводчиков

сгенерированные тексты?

Электронные библиотеки (примеры)

- Google Books - books.google.com
- Проект Гутенберг - www.gutenberg.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)
- Universal Digital [Library](https://www.digital-library.org)
- Internet [Archive.org](https://archive.org)

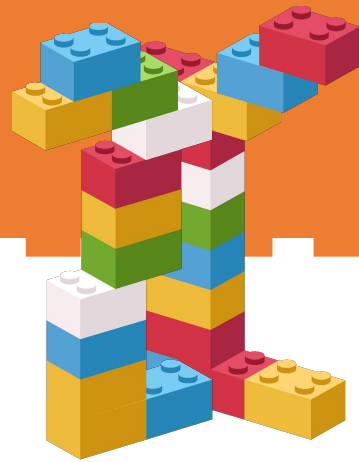


Сценарий 1 - документация языков

Корпус - лучше, чем текстовый архив

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

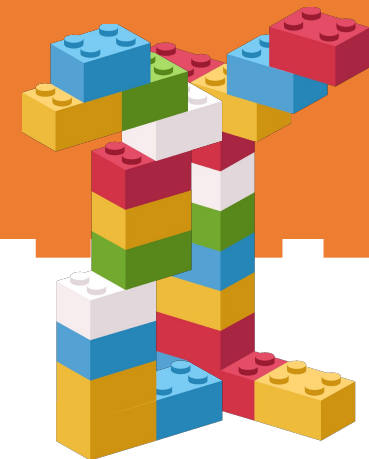
Электронные корпуса



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...)

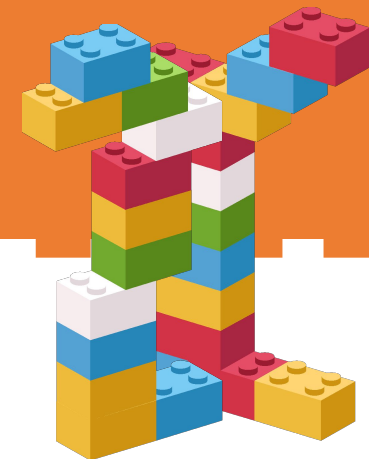
Электронные корпуса



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](https://www.gramota.ru))
 - другие специальные ресурсы (GIS-системы для диалектных исследований, [common voice](https://www.commonvoice.org/) 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 та или иная конструкция/единица
-
- если нет... можем спросить **других носителей!**
- собрать не прямые 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



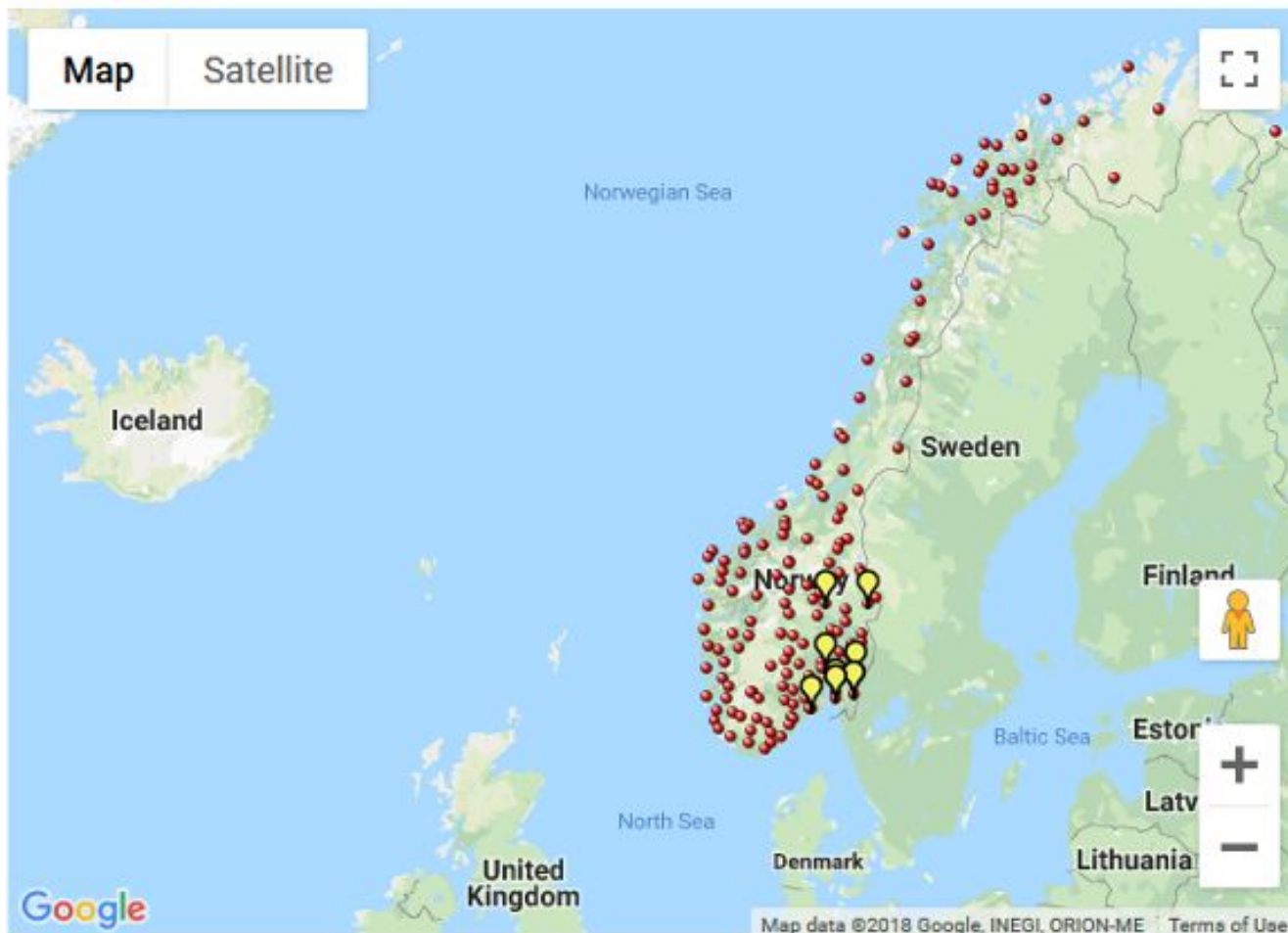
Корпуса

- 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
- Диалектные корпуса - HSE [LingConLab](#) corpora
- Учебные & эритажные корпуса - LINDSEI (Louvain International Database of Spoken English Interlanguage), REALEC





ekje ekk ekki ekkje **ennte ente** ett ette gge ijje ikj ikje ikk ikka ikke ikket ikki ikkj ikkje
ikkjee ikkji ikkke ikkkje ikø inggkje ingkji inngkje inngkji innkje innkji innt innte inntje inte issj issje
issjæ it itj itje itt itte itter itti ittj ittje ittjæ ittse ittse je ke kj kje kji kjæ kk kke kkj kkje kkji
nnte nte rrt rrtte 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

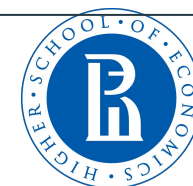
World Atlas of Language
Structures

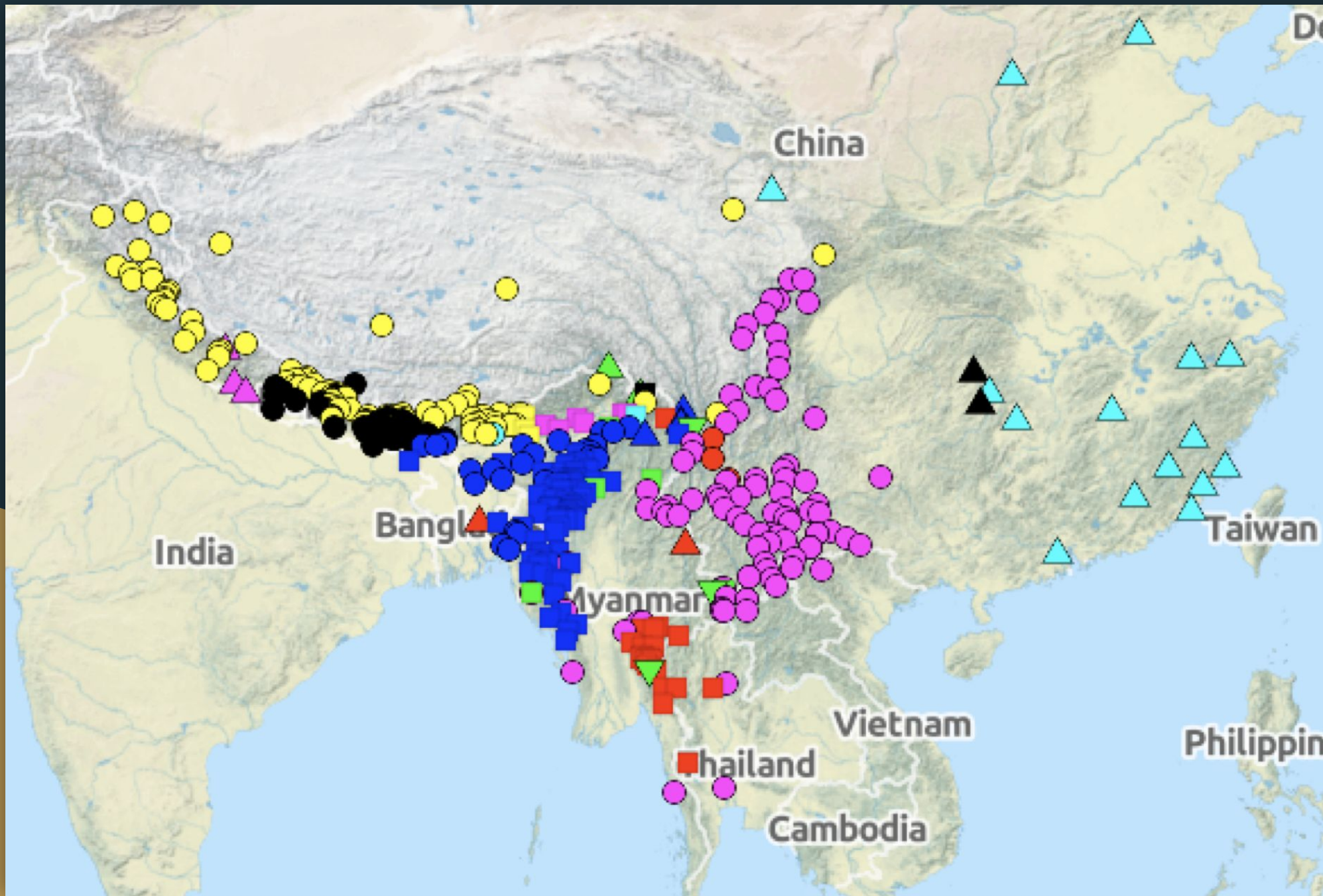
languages & languoids,
incl. sign languages and
artificial languages

Glottolog

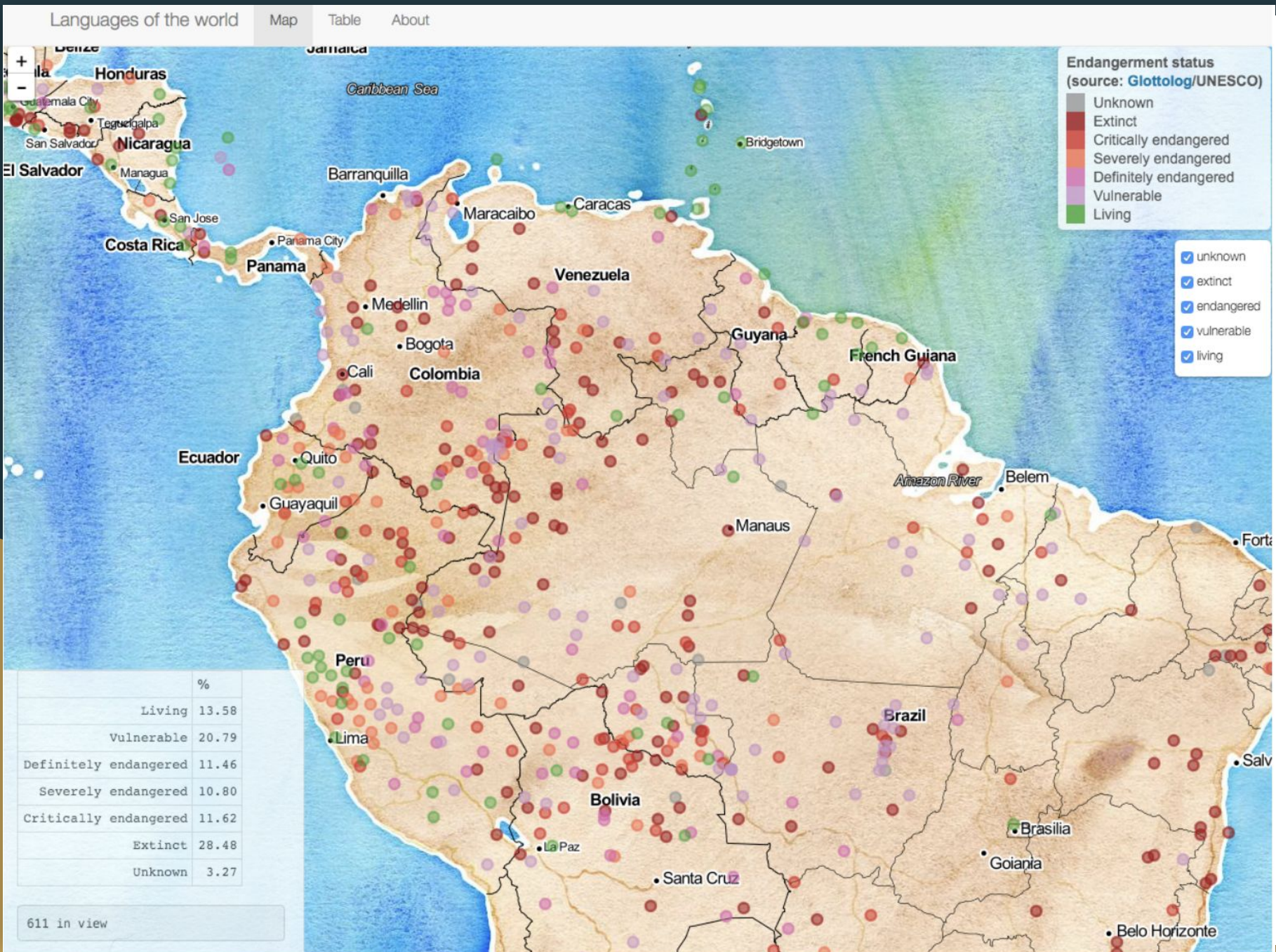
PHOIBLE

phonological
inventories



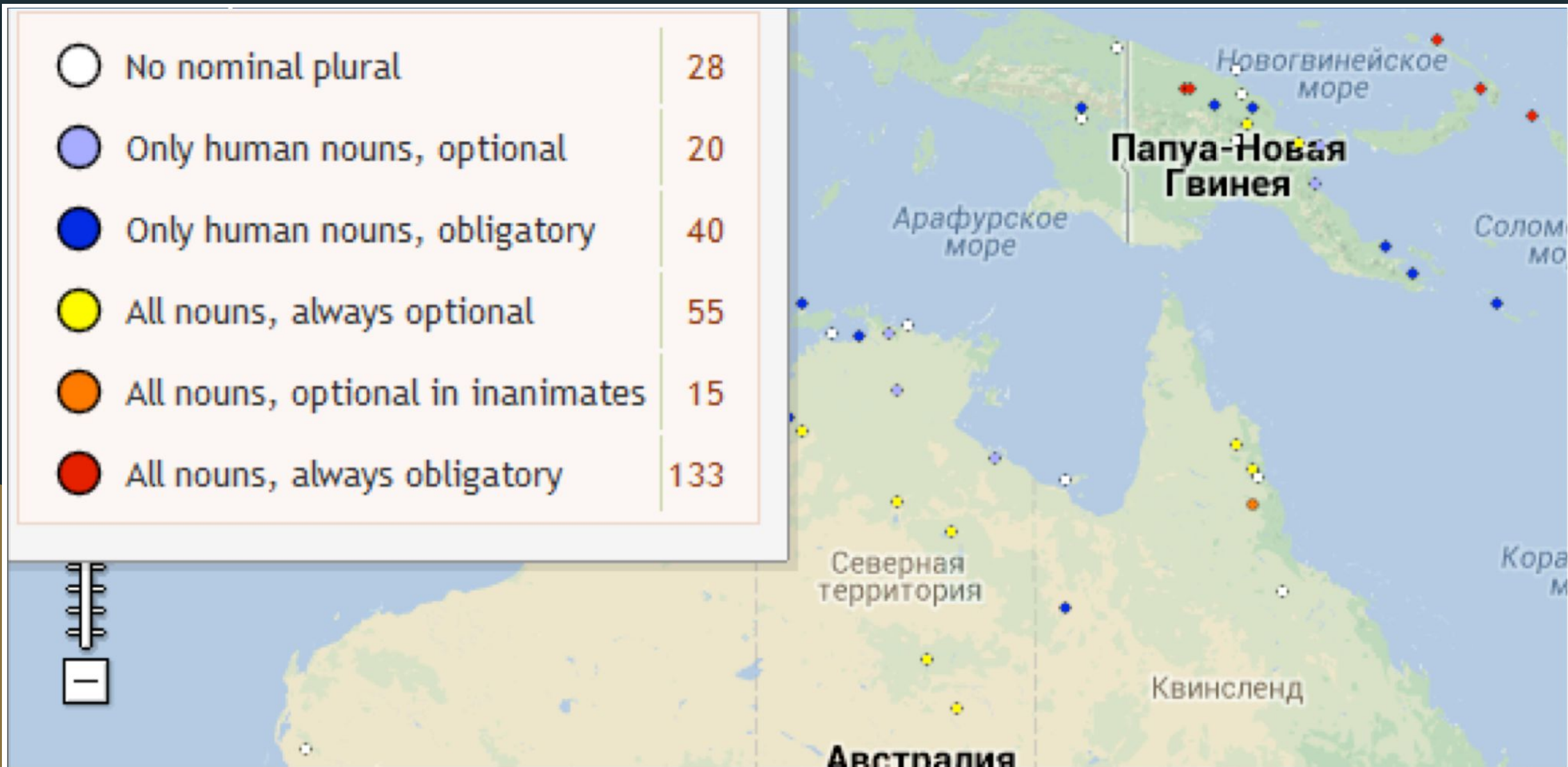


Glottolog:
Geographical
distribution of
Sino-Tibetan
languages



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

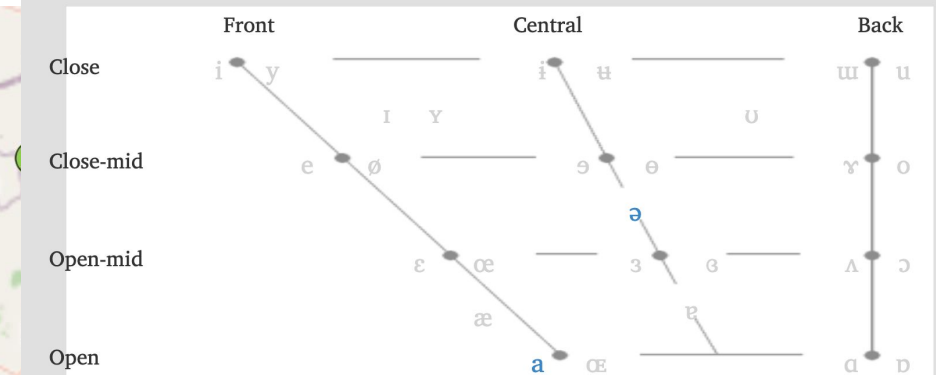


WALS: Feature 34A: Occurrence of Nominal Plurality

Consonants (Pulmonic)

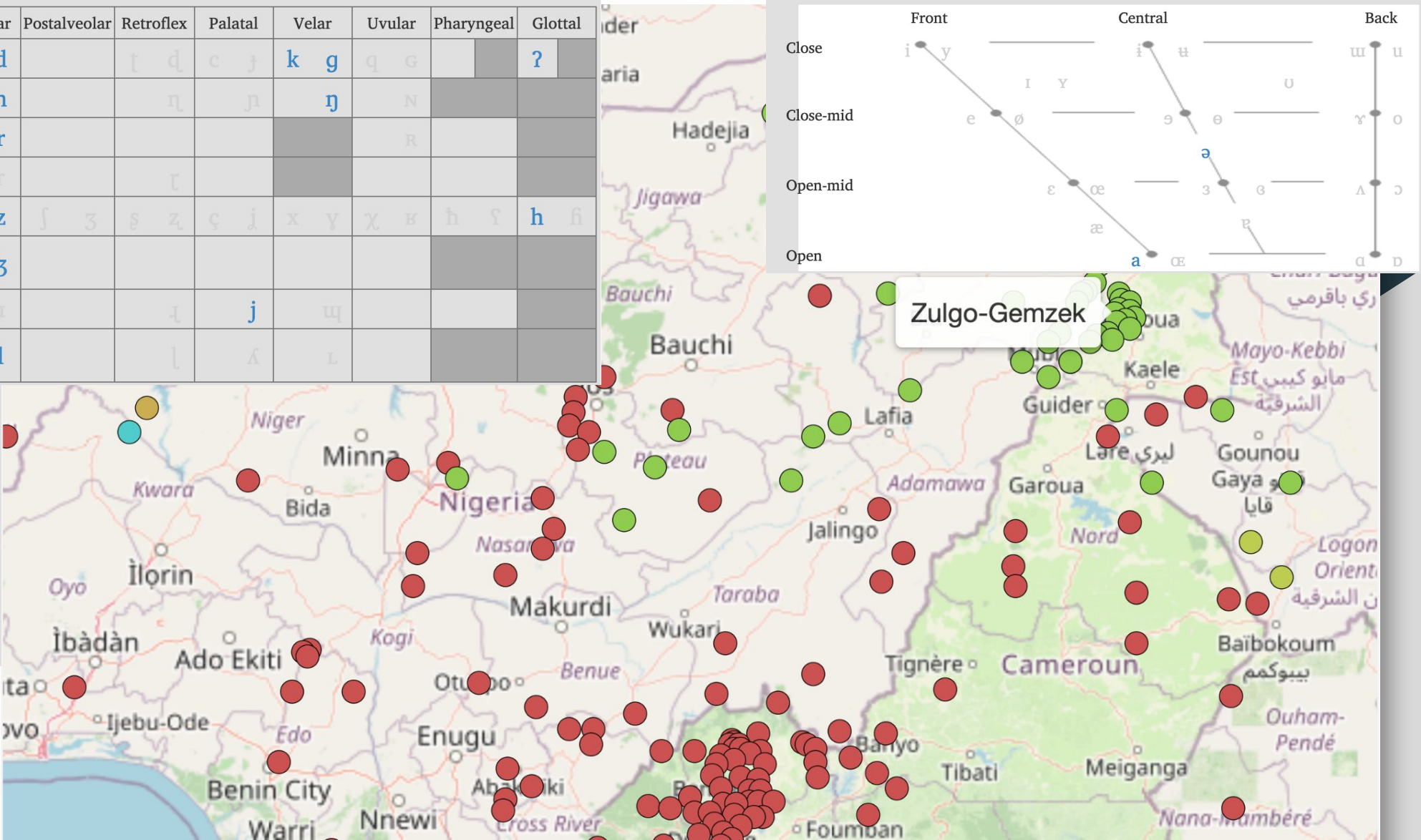
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t̪ d̪	t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ	n̪	n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ		r̪	r					ʀ		
Tap or Flap		ɸ	ɽ	ɾ		ɽ̌					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ			j	ɰ			
Lateral approximant			ɭ	l		ɭ̌	ʎ	ʟ			

Vowels



Consonants (Non-Pulmonic)

Clicks	Voiced implosives
ⱱ Bilabial	ɓ Bilabial
ⱦ Dental	ɗ Dental/alveolar
Ɱ (Post)alveolar	ɟ Palatal
Ɐ Palatoalveolar	ɠ Velar
Ɒ Alveolar lateral	ɡ Uvular

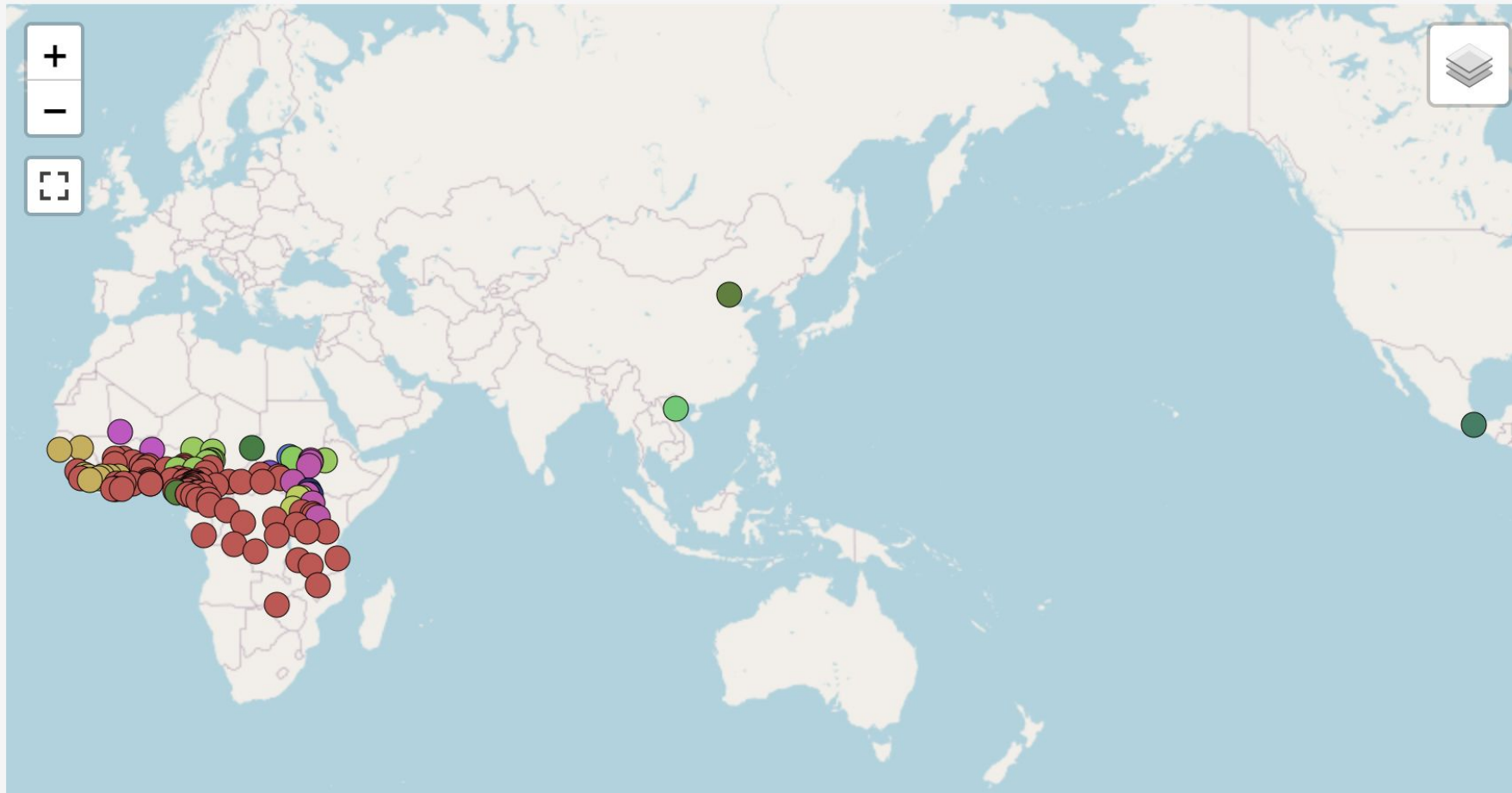


PHOIBLE: **zulgo** inventory: 2 vowels, 34 consonants

MODIFIER LETTER HIGH TONE BAR - MODIFIER LETTER LOW TONE BAR

Icon size ▾ ☐ Show/hide Labels

GeoJSON ▾



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

Базы данных

- Лексические

The Tower
of Babel

etymological database

VisuWords

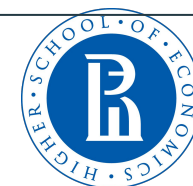
lexicon as a net:
visual dictionary,
visual thesaurus,
interactive lexicon

WordNet

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

Common
Voice

pronunciation
crowdsourcing
project



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), vï̯sse (T), vū̯ss (Kld.), vua̯ss (Not.) 'kleines Rentierkalb, bis es um den Peterstag neues Haar bekommt'

Mordovian: vaz (E M), vazné (E), vazná (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

FrameNet

Constructicon

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

- **Syntax, semantics, collocations**

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

Sketch Engine

Trados

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



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

<u>object of</u>	<u>58924</u>	<u>3.2</u>	<u>subject of</u>	<u>25451</u>	<u>2.4</u>	<u>modifier</u>	<u>67879</u>	<u>1.6</u>	<u>modifies</u>	<u>11026</u>	<u>0.3</u>
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

SketchEngine: word sketches of the noun GOAL

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
perceptive	0	34	0.0	6.4
thought-provoking	0	32	0.0	6.2
informed	0	66	0.0	6.2
autonomous	0	46	0.0	6.2
adaptive	0	39	0.0	6.1
well-informed	0	24	0.0	6.0
literate	0	26	0.0	5.9
compassionate	0	27	0.0	5.9
well-educated	0	17	0.0	5.7
cultured	0	19	0.0	5.7
rational	0	46	0.0	5.6
playful	0	22	0.0	5.6
sensitive	8	134	2.0	5.9
thoughtful	14	121	5.0	7.7
affectionate	6	31	4.5	6.2
sophisticated	23	75	4.2	5.7
charming	21	50	4.9	5.9
insightful	11	31	5.2	6.1
resourceful	12	29	5.8	6.3
witty	132	166	8.3	8.2
inventive	22	24	5.9	5.5
clever	54	30	5.8	4.8
funny	233	103	7.0	5.7
cunning	16	7	5.9	4.0
catchy	19	0	5.8	0.0

modifier	4950	3168	0.9	0.5
emotionally	0	111	0.0	8.6
artificially	0	52	0.0	7.9
fiercely	0	26	0.0	7.0
highly	0	570	0.0	6.9
ferociously	0	8	0.0	6.2
supposedly	0	28	0.0	6.2
averagely	0	7	0.0	6.1
moderately	0	11	0.0	5.7
reasonably	0	54	0.0	5.7
computationally	0	6	0.0	5.6
supremely	0	7	0.0	5.5
culturally	0	12	0.0	5.5
exceptionally	29	25	6.0	5.9
remarkably	24	11	5.7	4.8
amazingly	17	7	5.9	5.0
wonderfully	20	9	5.4	4.5
very	1707	596	5.6	4.0
too	476	76	5.4	2.8
damn	12	0	5.6	0.0
dead	16	0	5.8	0.0
diabolically	9	0	5.9	0.0
awfully	15	0	6.1	0.0
terribly	25	0	6.2	0.0
devilishly	17	0	6.8	0.0
fiendishly	45	0	8.1	0.0

modifies	10948	16081	2.0	2.4
being	0	208	0.0	6.1
robot	0	77	0.0	6.1
agent	9	455	0.4	6.0
guess	0	35	0.0	5.5
routing	0	27	0.0	5.3
layman	0	22	0.0	5.3
conversation	0	88	0.0	5.1
creature	11	137	2.4	5.9
lyric	81	80	5.8	5.7
fellow	52	14	5.1	3.1
pass	67	9	5.2	2.2
stuff	146	6	5.1	0.4
gimmick	15	0	5.1	0.0
satire	19	0	5.1	0.0
flick	21	0	5.2	0.0
lob	15	0	5.3	0.0
pun	19	0	5.3	0.0
ruse	17	0	5.5	0.0
eh	24	0	5.8	0.0
wordplay	21	0	5.8	0.0
chap	47	0	5.9	0.0
twist	94	0	6.5	0.0
trick	166	0	6.7	0.0
clog	50	0	7.0	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**



А есть ли у вас идеи?



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 исследованиям
- 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

