http://panlex.org

ËXTɾɛmE Localization: Translating Every Word in Every Language

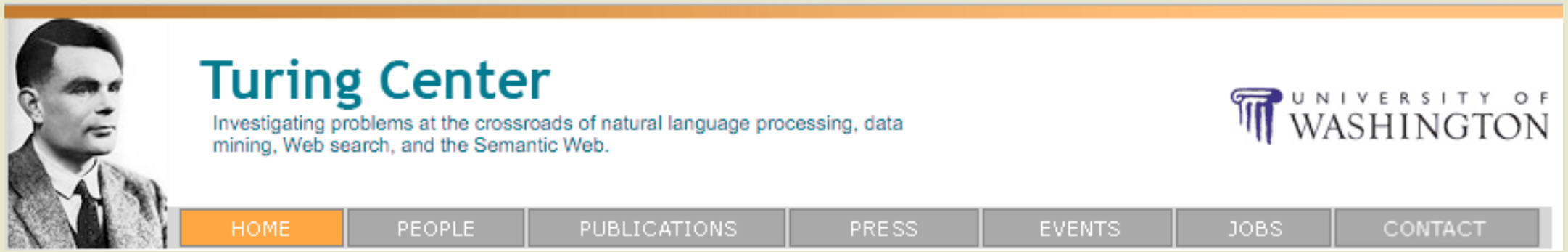
Outline

- Who we are
- Vision
- Strategy
- Concepts
- **Workflow**
- Results
- **Access**
- Help us
- Discussion

Who we are

The PanLex Project

- From 02005 to 02009 at:



- From 02010 to 02011 at:

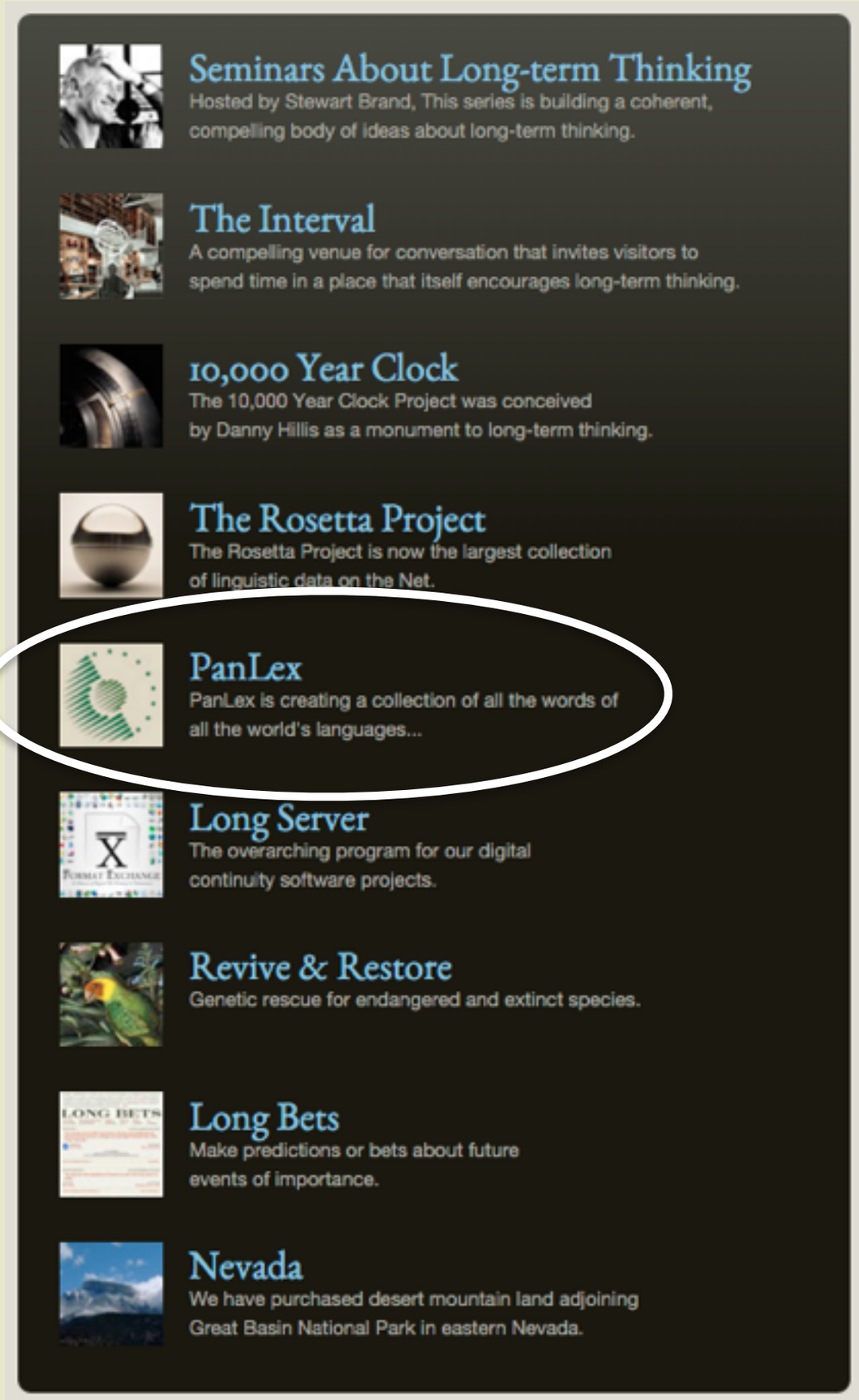
Utilika Foundation

Who we are

- From 02012 to now at:

THE LONG NOW
FOUNDATION

http://longnow.org



Who we are

- A partner or member of:
  - The Rosetta Project
  - Internet Archive
  - Unicode Consortium
  - Endangered Languages Project
  - CJK Dictionary Institute
  - Kamusi
  - Open Knowledge Foundation
  - Universität Leipzig
  - and others

Who we are

- –02010:

Oren Etzioni	Marcus Sammer	Daniel S. Weld
Mausam	Timothy Baldwin	Jeff Bilmes
Stephen Soderland	Katherine Everitt	Katrin Kirchhoff
Kobi Reiter	Christopher Lim	Bo Qin
Michael Skinner	Janara Christensen	

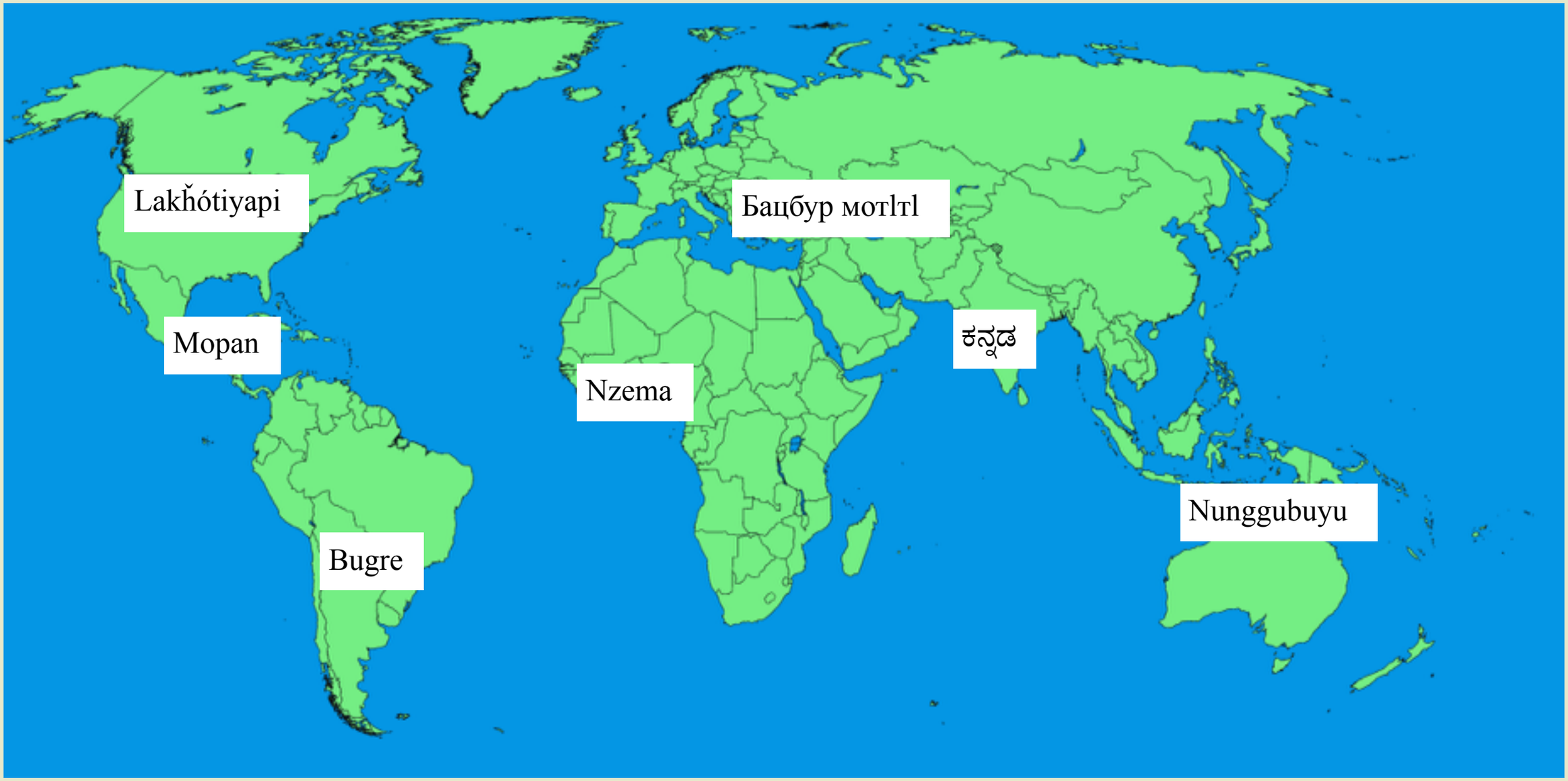
- 02010–now (* = here at this talk):

Jonathan Pool*	Alex DelPriore*	Julie Anderson*
David Kamholz*	Gary Krug*	Emily Bender
Susan Colowick	Benjamin Yang*	Steven Bird

- And interns, volunteers, advisors

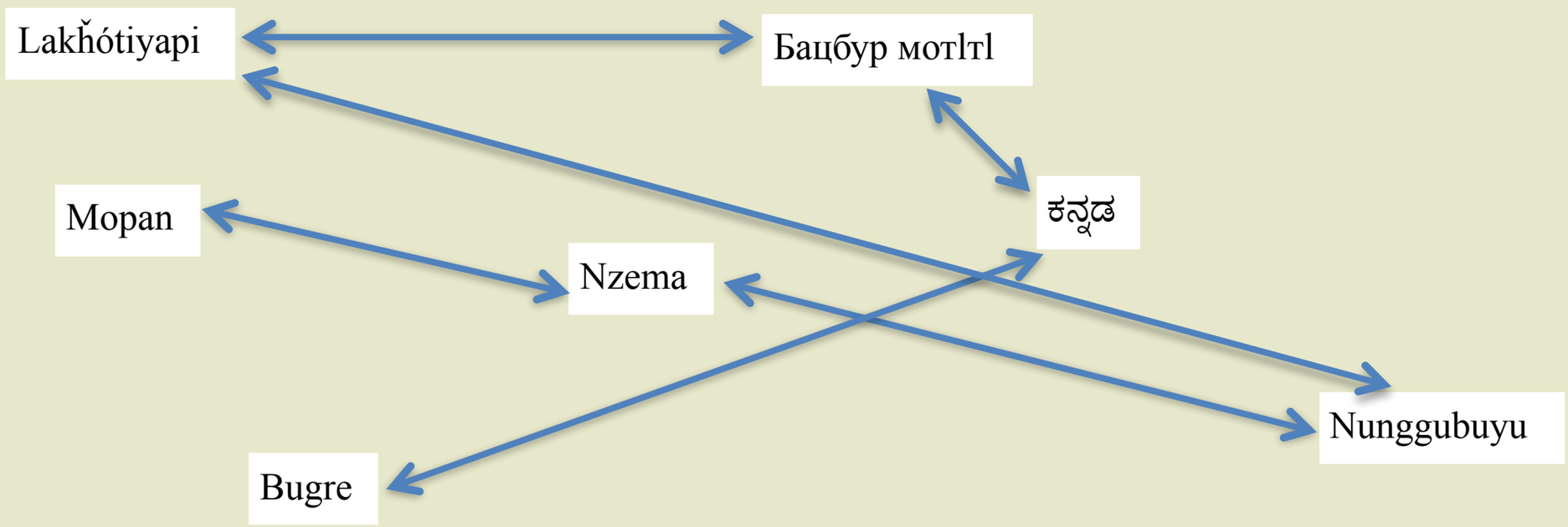
Vision

- Every language a world language.



Vision

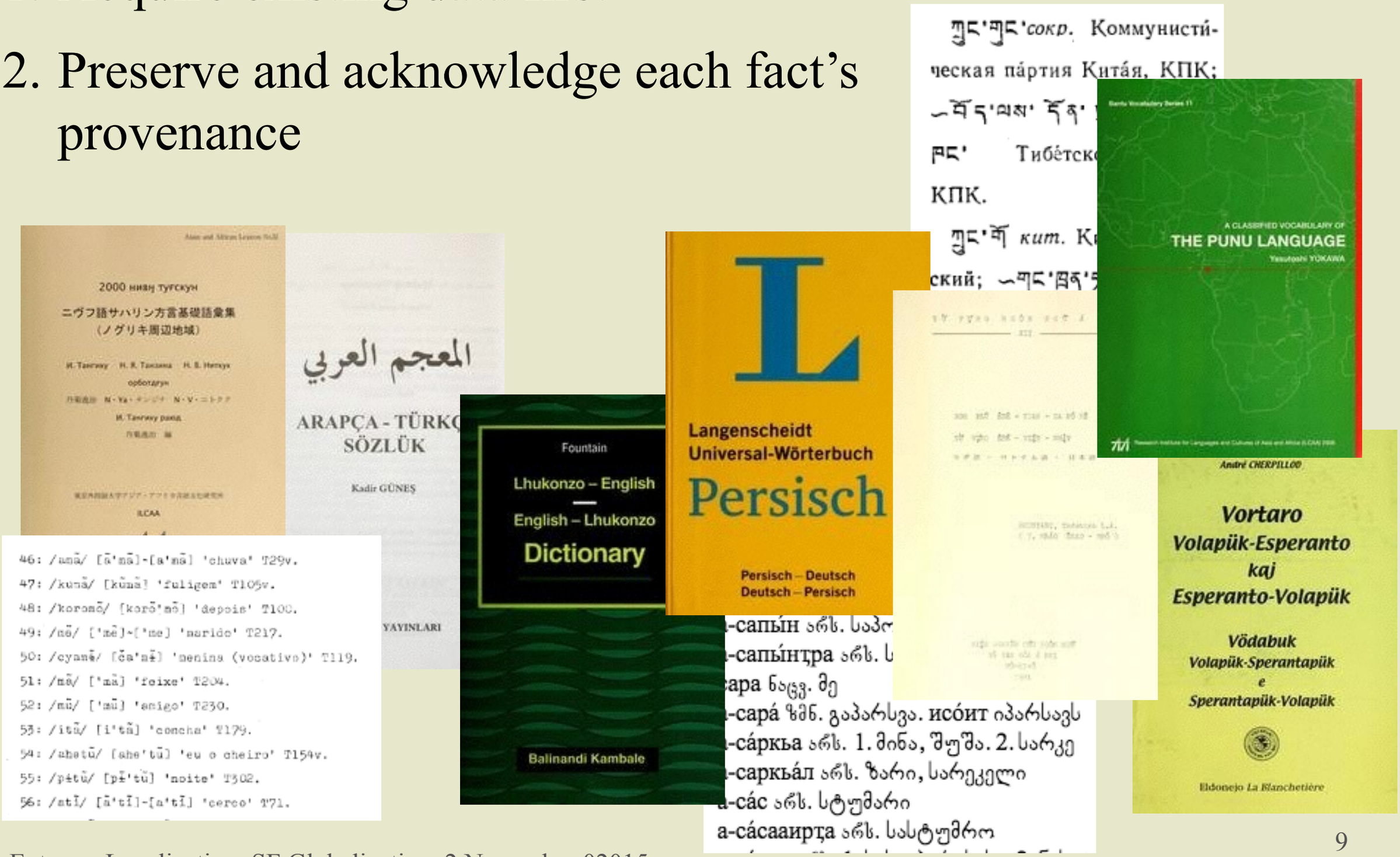
- Any language translatable into any other language.



≅ 7K languages ⇒ 50M pairs

Strategy

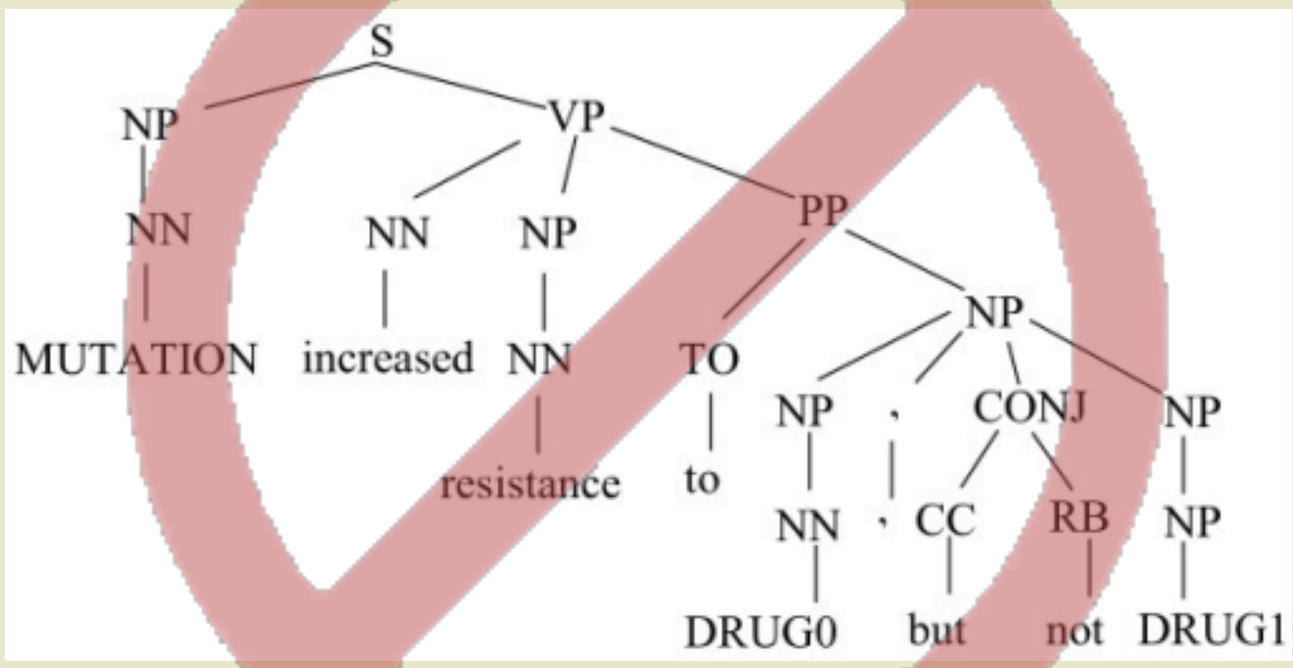
1. Acquire existing data first
2. Preserve and acknowledge each fact's provenance



Strategy

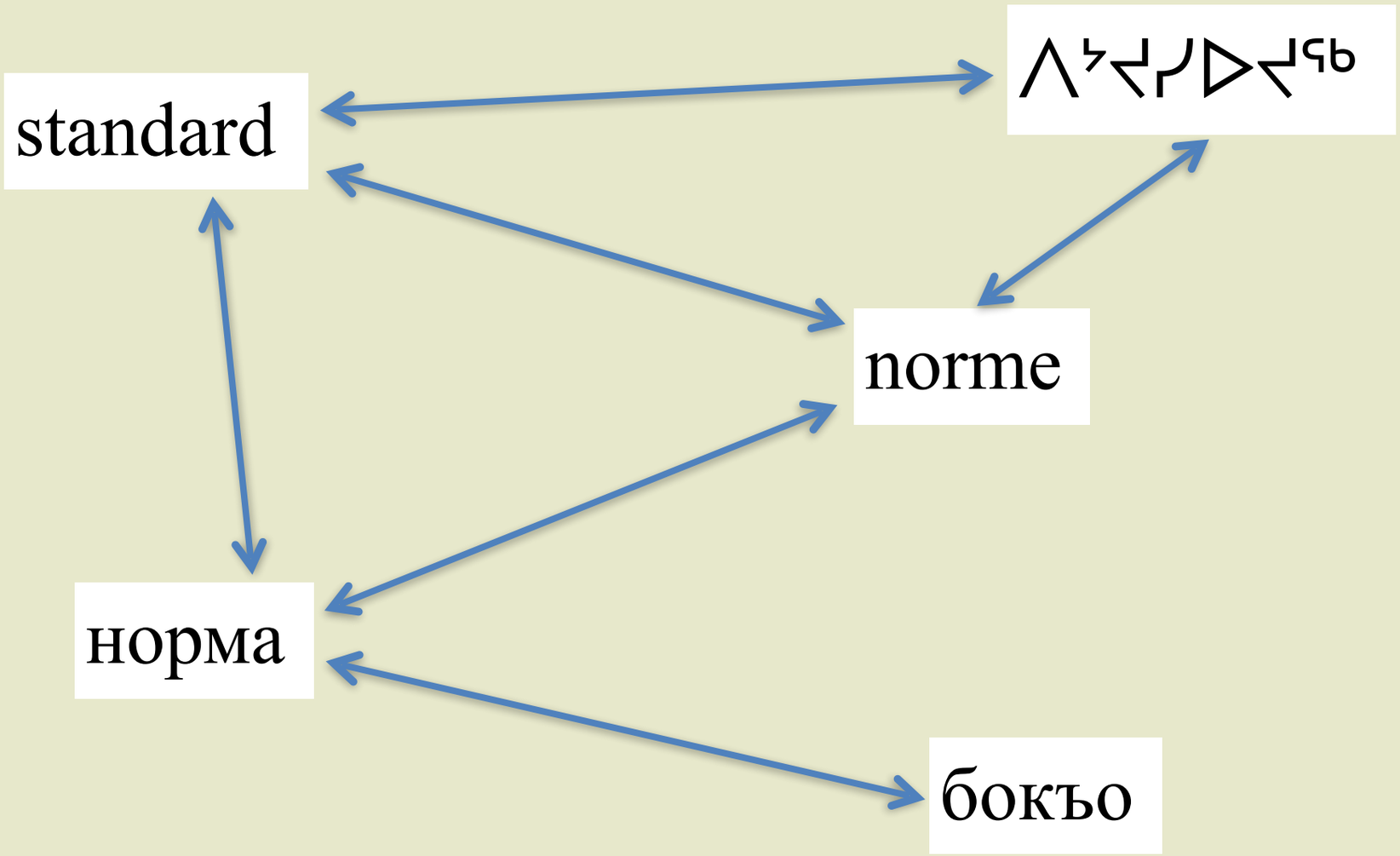
3. Start with lexicon, defer grammar

17596085 qaq
17605846 qaritte=ma
17603016 qaro
17369810 qaru
17109503 qarun-kó
17577214 qarunko
17577239 qarunte
17585713 qaw
17222182 qaw-
17093499 qaw-xó
17222324 qawai
17095202 qawho
17605754 qawtitto
17221018 qaww-
17093643 qawxó
17342725 qawčitto
17606812 qayanko
17575892 qayanqayo
17607397 qayanɠayo



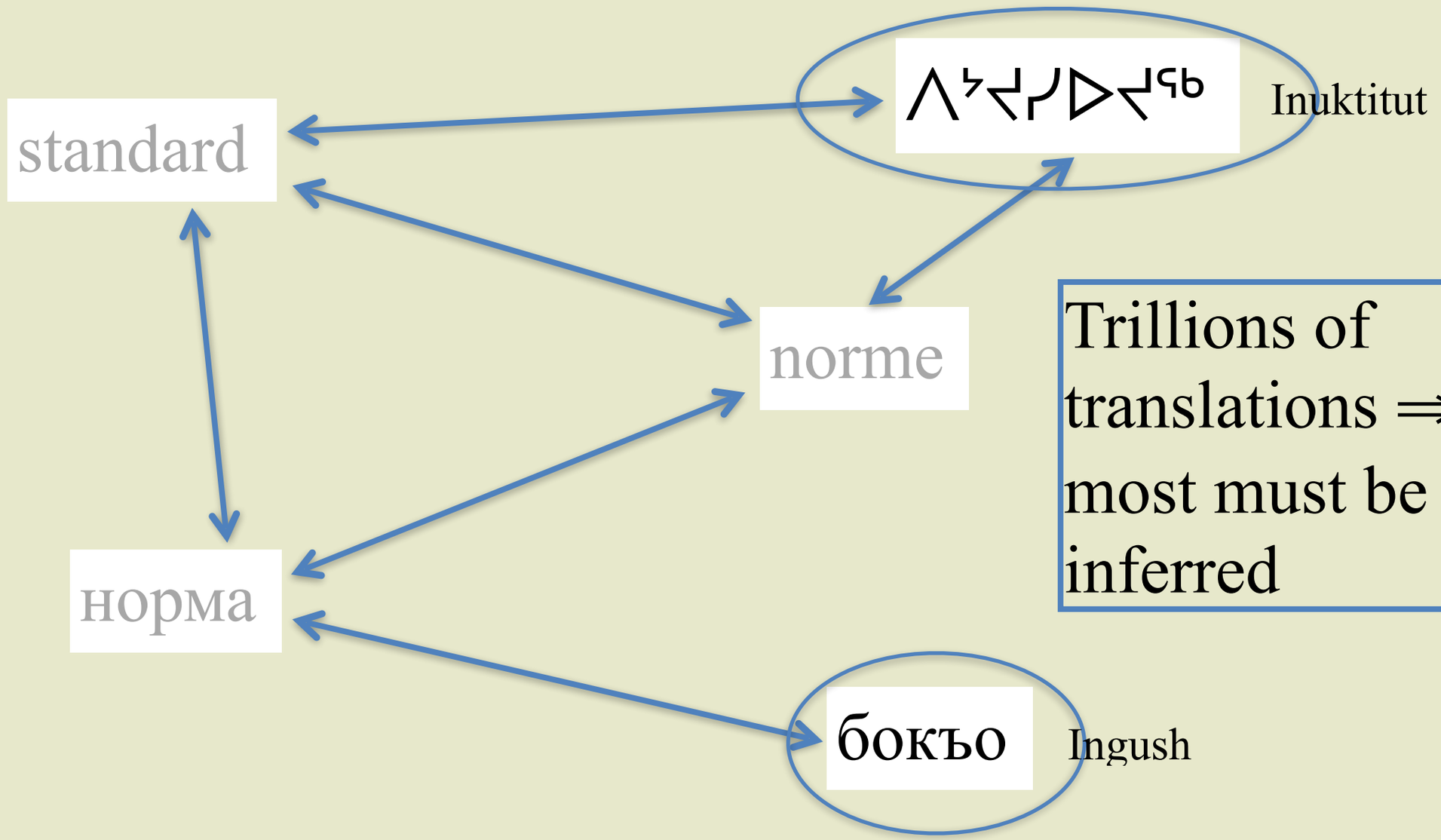
Strategy

4. Translate:
   - Lexemes
   - Symmetrically
   - Any-to-any



Strategy

5. Infer missing translations:



Concepts

Languages have ***varieties***.

azj-000	azərbaycanca
azj-001	Азәрбајчан дили
azj-002	Tərəkəməcə
azj-003	терекеме
azj-004	آذربایجانی

Language varieties have ***expressions***.

18433157	ишдә
18434129	ишдән бојун гачырмаг
18434128	ишдән гачмаг
18438369	иши чох олан
18434163	ишләдилмиш
18434157	ишләдилмәк
18453099	ишләмә

Concepts

Expressions have ***meanings***.

Expressions sharing a meaning are ***distance-1 translations*** (including synonyms) of each other.

source — 3081 — pum-eng-npi:Rai

meaning — 19280118

denotation					more
code	language		expression		
48076120	npi-000	नेपाली	2604048	छुचुन्द्रो	
48076119	pum-000	Pumā	17568478	cʌktoŋ	

Each meaning has a ***source***.

The assignment of a meaning to an expression is a ***denotation***.

source — 372 — npi-epo:Mandahar

meaning — 4519169

denotation					more
code	language		expression		
13528666	epo-000	Esperanto	650223	talpo	
13528665	npi-000	नेपाली	2604048	छुचुन्द्रो	

Concepts

Definitions are distinct from expressions.

`rus-000`: поздно закрывающийся или круглосуточный магазин

`eng-000`: convenience store

`jpn-000`: コンビニ

Workflow

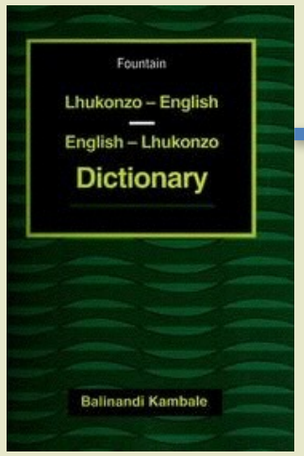
- Source acquisition
  - Discovery
  - Procurement
  - Registration

number	4465
PanLex — beginning	2014-02-26
name	lem-fra:Taylor
World Wide Web	http://www.sil.org/resources/archives/47934
ISBN	
author	Carolyn Taylor; Terri R. Scruggs
title	Lexique nɔmaándɛ́ - français
publisher	SIL Cameroon
year	2003
good — number	6
source — primary	4465
other	SIL ID 47934; yr:2003
permission — kind	cc
right — text	Unless stated otherwise in the file or the item description, all items are available under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported License.
right — person — name	
permission — email — address	

Ɔ - ɔ

ɔbaca *n.* bouton de gale.
ɔbála *n.* chaleur, fièvre.
ɔbalasaná *n.* sorte de termite qui apparaît en nov. (à partir de 16h.).
ɔbɛ́lɛ *n.* trou de fourmis.
ɔbɛ́lɛ *n.* trou de sortie des termites.
ɔbɔbɛ *n.* charme.

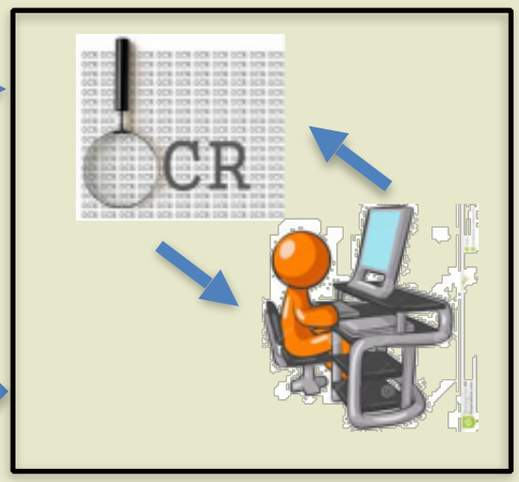
Workflow

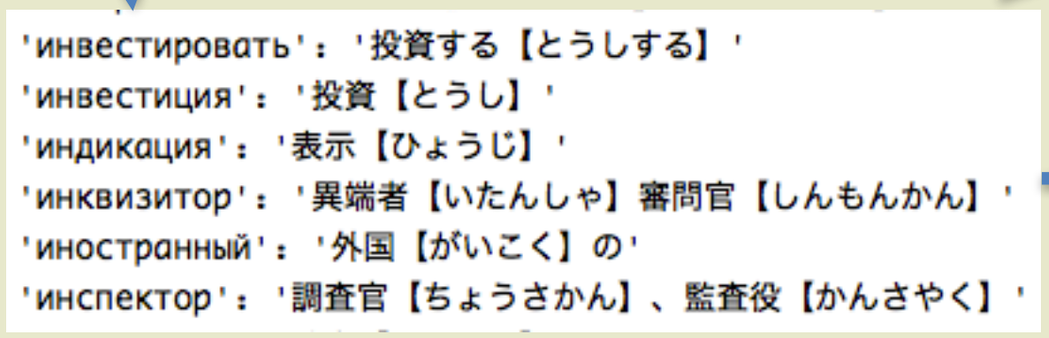


Book

Internet Archive

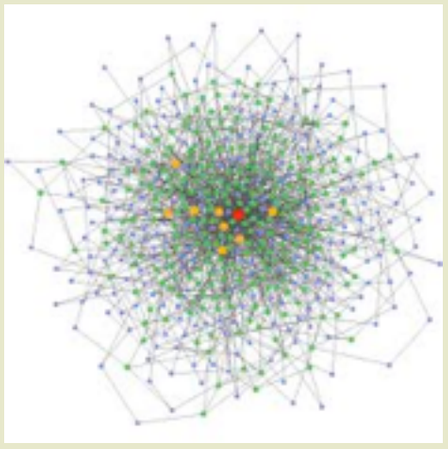
Image file





Text file

PanLex Source Analysts



PanLex

Workflow

- Text conversion (if necessary):
  - Recognition (OCR)
  - Human (selected) text entry
  - Recoding (anything to Unicode)
  - Character normalization (NFC etc.)
  - Standardization (diacritics, punctuation, letter case, etc.)

Workflow

- Tabularization:

balima `(YR, YY)` *noun*

1 heaven, sky camp. A specific spot in the sky world. *Gunagala* is more like the English 'heaven'.

2 place far away. This is a euphemism for 'as far from human affairs as possible'.

Also 'bulima'.

balima	noun	heaven▸sky camp
balima	noun	place far away

Workflow

- Serialization:

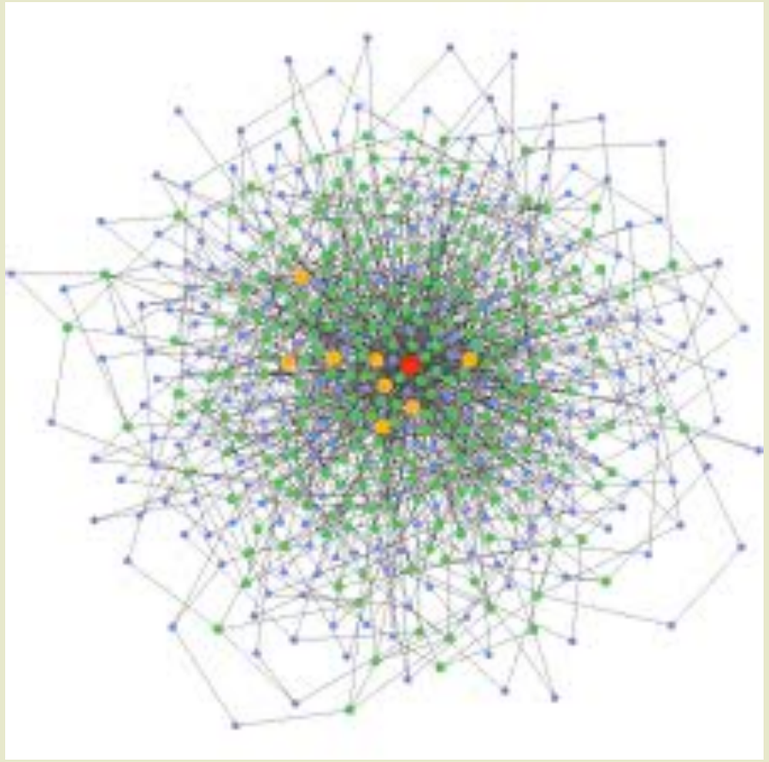
balima	noun	heaven▸sky camp
balima	noun	place far away

mn
dn
kld-000
balima
dcs
art-303
PartOfSpeechProperty
art-303
Noun
dn
eng-000
heaven
dn
eng-000
sky camp

mn
dn
kld-000
balima
dcs
art-303
PartOfSpeechProperty
art-303
Noun
df
eng-000
place far away

Workflow

- Validation and submission:



mn
dn
kld-000
balima
dcs
art-303
PartOfSpeechProperty
art-303
Noun
dn
eng-000
heaven
dn
eng-000
sky camp

mn
dn
kld-000
balima
dcs
art-303
PartOfSpeechProperty
art-303
Noun
df
eng-000
place far away

Workflow

- Three case studies:
  1. Ben Yang: Cherokee
  2. Gary Krug: Spanish dialects
  3. Alex DelPriore: South Sámi

Case 1. Ben Yang: Cherokee

Source selection

stop
beginning
back

PanLex 4.0

[PanLem (aprsel1w)]
[9.4.5]

you — benayang

source — edit — fact — see

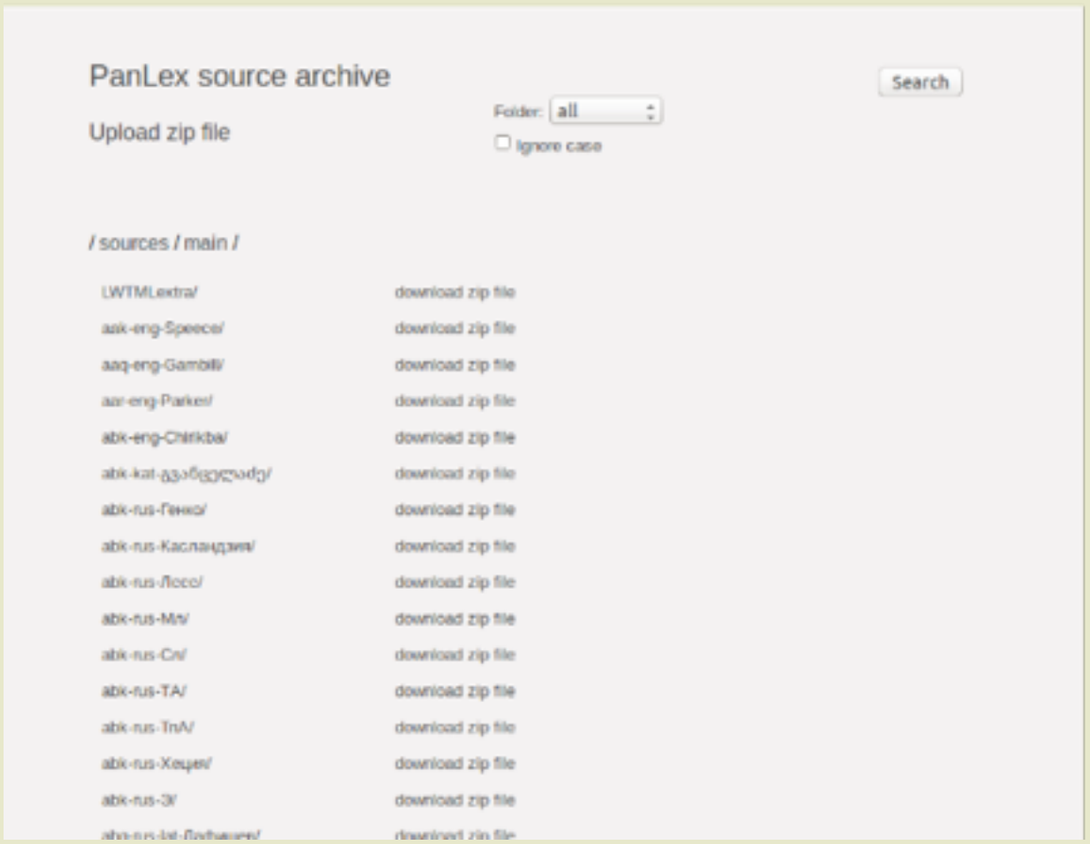
edit — done — no

see	source	difficult	publisher	World Wide Web	language
	aaq-eng:Gambill	4	Freelang	http://www.freelang.net/dictionary/abena	aaq-000[1478]; eng-000[2946311]
	aar-eng:Parker	6	Dunwoody Press	http://www.dunwoodypress.com/products/-/	aar-000[1725]; eng-000[2946311]
	abk-eng:Chirikba	8		http://apsnyteka.org/874-chirikba_a_dict	abk-002[181]; abk-003[14]; abk-004[1]; abk-005[1]; abk-006[275]; abk-007[1]; abk-008[0]; abk-009[1]; abk-010[1]; eng-000[2946311]
	abk-rus:Генко	8	Издательство “Алашар	http://apsnyteka.org/310-genko_abkhazsko	abk-000[10990]; rus-000[1200755]
	abk-rus:Каслазиа	4	ОЛМА-Пресс	http://www.abiblioteka.info/rus/lang/sec	abk-000[10990]; rus-000[1200755]
	abq-rus-lat:Лафишев	7	Карачаево-Черкесское	http://abazinka.ru/books.html	abq-000[11927]; lat-000[84180]; rus-000[1200755]
	abq-rus:Адзинов	8	Издателство “Советск	http://abazinka.ru/books.html	abq-000[11927]; rus-000[1200755]
	abq-rus:Кълыч	7	Карачаево-Черкесское	http://abazinka.ru/books.html	abq-000[11927]; rus-000[1200755]
	abq:ТобылъТЩ	7	Къарча-Черкес Респуб		abq-002[1]; abq-003[1]
	abq:ТобылъЩТ	7	Къарча-Черкес Респуб		abq-002[1]; abq-003[1]
	abt-tpi-eng:Ambia	7	SIL International	http://www.sil.org/resources/archives/31	abt-005[1]; eng-000[2946311]; tpi-000[3316]
	abt-tpi-eng:Bakandu	7	SIL International	http://www.sil.org/resources/archives/31	abt-006[1]; eng-000[2946311]; tpi-000[3316]
	abt-tpi-eng:Kerry	6	SIL International	http://www.sil.org/resources/archives/31	abt-004[1]; eng-000[2946311]; tpi-000[3316]
	abt-tpi-eng:Kundama	4	Summer Institute of	http://www.sil.org/pacific/png/show_work	abt-000[195]; eng-000[2946311]; tpi-000[3316]
	abt-tpi-eng:Wilson	7	SIL International	http://www.sil.org/resources/archives/31	abt-003[38]; eng-000[2946311]; tpi-000[3316]
	ach-eng:Blackings	6	LINCOM		ach-000[112]; eng-000[2946311]
	acu-spa:Fast	7	Ministerio de Educac	http://www.sil.org/resources/archives/29	acu-000[675]; spa-000[540026]
	acw-eng:Omar	7	Foreign Service Inst	https://archive.org/details/Fsi-SaudiAra	acw-001[1]; eng-000[2946311]
	adx-jpn:海老原	8	アジア・アフリカ言語文化研究所	http://www.aa.tufs.ac.jp/documents/train	adx-000[1]; jpn-000[505498]

Case 1. Ben Yang: Cherokee

Data retrieval

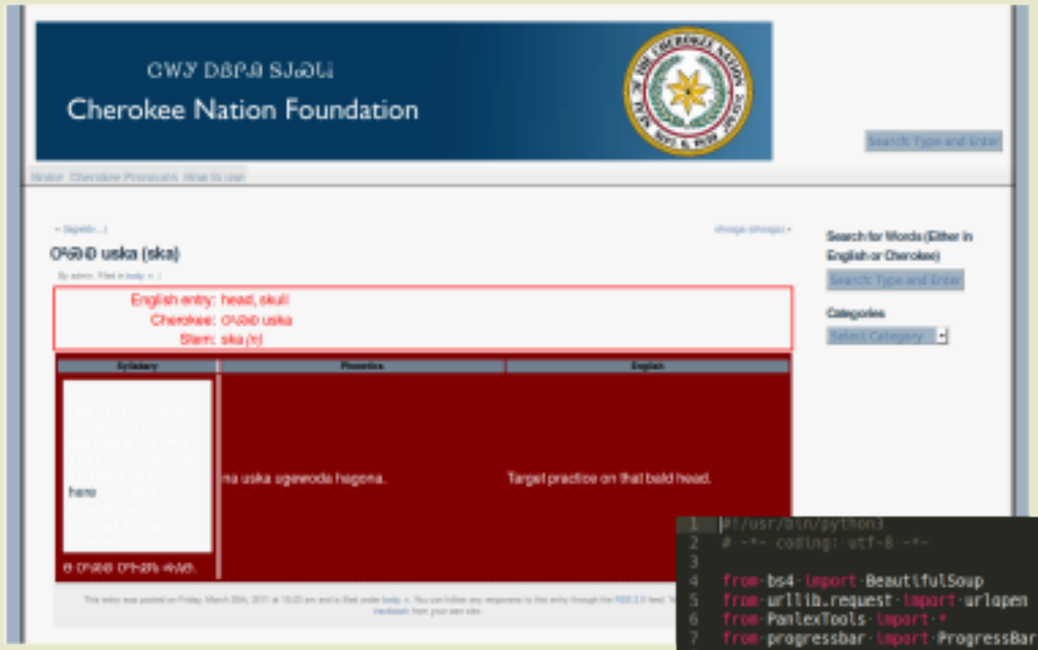
Data retrieved and archived during source acquisition

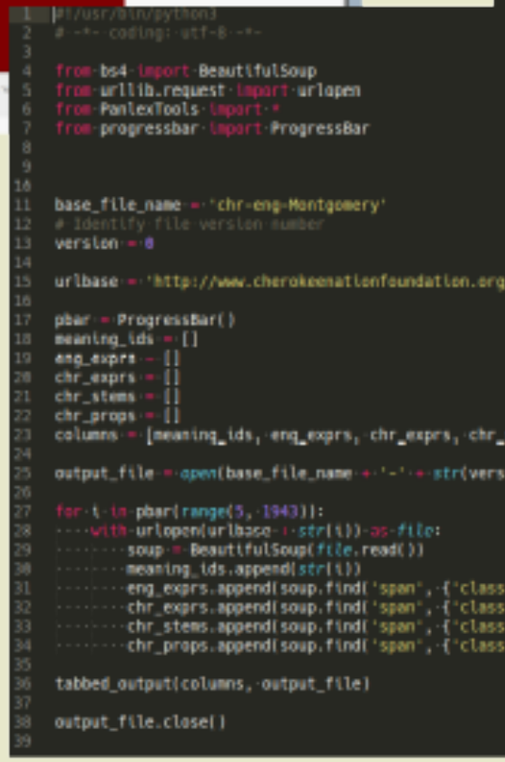


Some sources are living sources

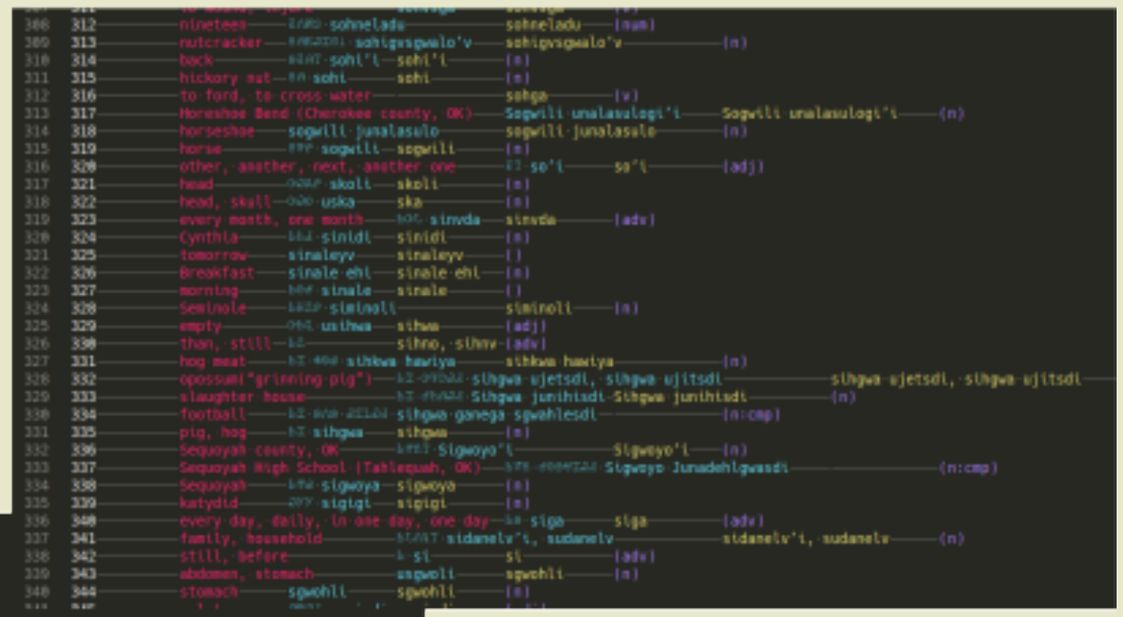
Case 1. Ben Yang: Cherokee

Data extraction





```
#!/usr/bin/python3
# -*- coding: utf-8 -*-

from bs4 import BeautifulSoup
from urllib.request import urlopen
from PanlexTools import *
from progressbar import ProgressBar



base_file_name = 'chr-eng-Montgomery'
# Identify file version number
version = 0

urlbase = 'http://www.cherokeenationfoundation.org/dictionary/?p='

pbar = ProgressBar()
meaning_ids = []
eng_exprs = []
chr_exprs = []
chr_stems = []
chr_props = []
columns = [meaning_ids, eng_exprs, chr_exprs, chr_stems, chr_props]

output_file = open(base_file_name + '-' + str(version) + '.txt', 'w', encoding='utf-8')

for i in pbar(range(5, 1043)):
    with urlopen(urlbase + str(i)) as file:
        soup = BeautifulSoup(file.read())
        meaning_ids.append(str(i))
        eng_exprs.append(soup.find('span', {'class' : 'glo'}).text)
        chr_exprs.append(soup.find('span', {'class' : 'def'}).text)
        chr_stems.append(soup.find('span', {'class' : 'stm'}).text)
        chr_props.append(soup.find('span', {'class' : 'partsp'}).text)

tabbed_output(columns, output_file)

output_file.close()
```

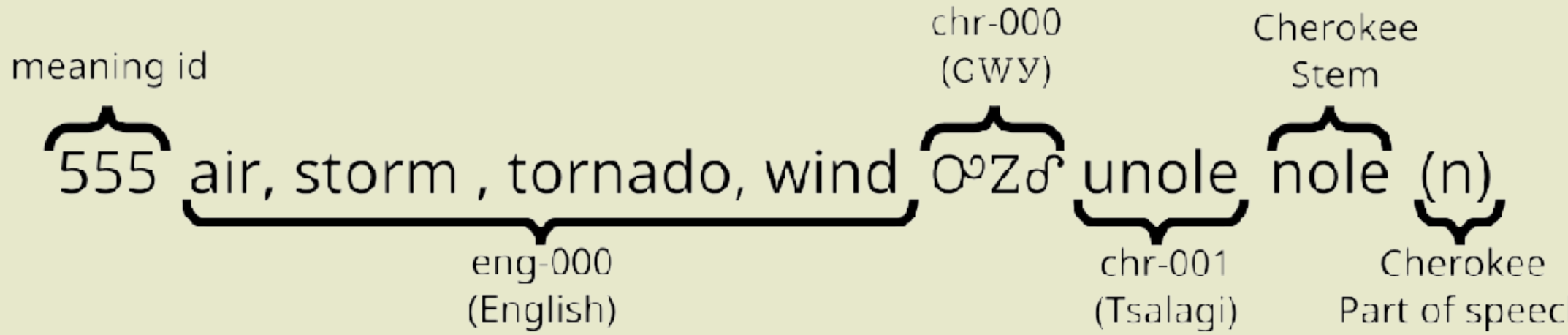


312	nineteen	ᎠᏂᎦᏓ sohnelada	sohneladu	[num]	
313	nutcracker	ᏍᎩᏳᏍᏘᏗᏍᏓᏩᏗᏙ sohigvsgwalo'v	sohigvsgwalo'v	[n]	
314	back	ᏍᎻᎵ sohl'i	sohl'i	[n]	
315	hickory nut	ᏍᏂ sohi	sohi	[n]	
316	to ford, to cross water		sohga	[v]	
317	Horseshoe Bend (Cherokee county, OK)	Sogwili unalasulogi'i	Sogwili unalasulogi'i	[n]	
318	horseshoe	sogwili junalasulo	sogwili junalasulo	[n]	
319	horse	ᏍᏆᏓᎵ sogwili	sogwili	[n]	
320	other, another, next, another one		so'i	so'i	[adj]
321	head	ᎤᏍᏆᏗ skoli	skoli	[n]	
322	head, skull	ᎤᏍᏆᏗ uska	ska	[n]	
323	every month, one month	ᏰᎥᏃ sinvda	sinvda	[adv]	
324	Cynthia	ᏐᏂᏆ sinidi	sinidi	[n]	
325	tomorrow	sinaleyv	sinaleyv	[]	
326	Breakfast	sinale ehi	sinale ehi	[n]	
327	morning	ᏐᎾᎵ sinale	sinale	[]	
328	Seminole	ᏐᎴᏁᎵ siminoli	siminoli	[n]	
329	empty	ᏍᏂᏫ usihwa	sihwa	[adj]	
330	than, still	Ꮟ	sihno, sihnv	[adv]	
331	hog meat	ᏏᎩᏯ sihkwa hawiya	sihkwa hawiya	[n]	
332	opossum("grinning pig")	ᏏᏅᏍᏆᏗ sihgwa ujetsdi, sihgwa ujitsdi	sihgwa ujetsdi, sihgwa ujitsdi		
333	slaughter house	ᏏᎩᏯᏂᏍᏆᏗ Sihgwa junihisdi	Sihgwa junihisdi	[n]	
334	football	ᏏᎩᏯ ᎦᏂᎦᏗᏓᏁᏗ sihgwa ganega sgwahlesdi		[n:cmp]	
335	pig, hog	ᏏᎩᏯ sihgwa	sihgwa	[n]	
336	Sequoyah county, OK	ᏍᎩᏯᏱᎩ Sigwoyo'i	Sigwoyo'i	[n]	
337	Sequoyah High School (Tahlequah, OK)	ᏍᎩᏯᏱᏂᏍᏆᏗ Sigwoyo Junadehlgwasdi		[n:cmp]	
338	Sequoyah	ᏍᎩᏯᏱ sigwoya	sigwoyo	[n]	
339	katydid	ᏍᏆᏀᏆᏗ sigigi	sigigi	[n]	
340	every day, daily, in one day, one day	ᏍᎦ siga	siga	[adv]	
341	family, household	ᏍᏗᏓᏂᏙᏱ sidanelv'i, sudanelv	sidanelv'i, sudanelv	[n]	
342	still, before	Ꮟ si	si	[adv]	
343	abdomen, stomach		sgwoli	sgwohli	[n]
344	stomach	sgwohli	sgwohli	[n]	

Case 1. Ben Yang: Cherokee

Tabularization

- Organize data



- Fix regular errors
  - t → Ꮭ
- Separate synonyms
  - “human, person, people” → human‣person‣people
- Lemmatize
  - “to boil (T)” → boil (verb-transitive)

Case 1. Ben Yang: Cherokee

Serialization

- Regularization of data
  - code point fixing
  - standardization of properties and classifications

```
verb suffix art-303:Morpheme:art-303:Suffix
verb-intransitive art-303:PartOfSpeechProperty:art-303:IntransitiveVerb
verb-transitive art-303:PartOfSpeechProperty:art-303:TransitiveVerb
verbs art-303:PartOfSpeechProperty:art-303:Verbal
```

- Normalization
- Output

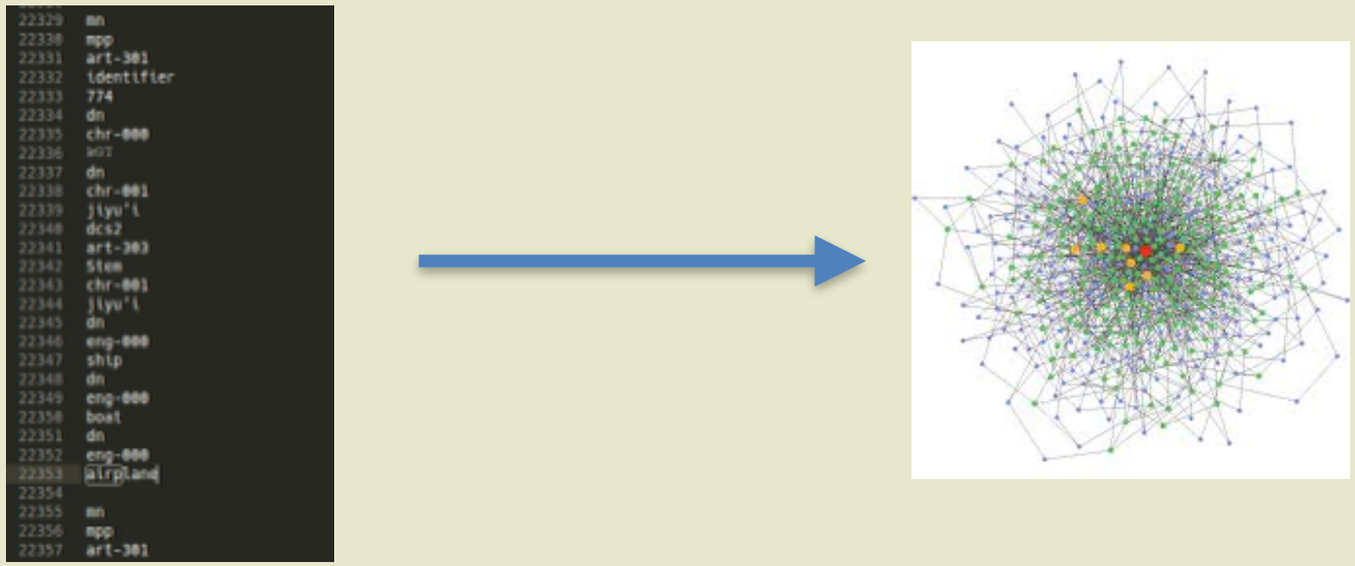
```
773-ᎯᎷᎠ jiyukdi-jiyukdi-(adv) straight
774-ᎯᏀᏆ-jiyu'i jiyu'i ()-ship•boat•airplane
775 jiyu junidisdi jiyu junidisdi () airport
776-ᎯᏀ-ᏗᎠᏓᏍᏙ jiyu ahlawidisgi (n:cmp) plane
777 ᎯᏀ jiyu jiyu (n) boat•canoe•plane
778-ᎤᏂᏯ-jiya jiya (n)-worm
779-ᎯᏯ jiya jiya (n)-otter
780-ᎯᏤᎦ-jitaga jitaga (n)-chicken
781 jislvsga jislvsga (v)-light a fire•light a match VERB
782-ᎯᏍᎩᏯᏍᏗ-jisgwalisda-jisgwalisda-(n)-blackbird
783-ᎯᏍᎩ-jisgwa jisgwa (n) bird
784-ᎠᏂᏴᏫᏍᏗᏢᎢ Jisgvnigesdv'i Jisgvnigesdv'i (n) Gore (place of carp)
785 ᎯᏍᎴᏁ jisdvna jisdvna (n) crawfish•crayfish•crawdad
786-ᎯᏍᏕ-jisdu•jisdi-jisdu•jisdi-(n)-rabbit
787-ᎯᎷ jimi jimi (n)-Jimmy
788 jilvsgi-jilvsgi-(n)-flower
789 degatsilugv (CL) jihlusga (v)-squat VERB
790-ᎯᎩ jigi jigi (v)-be (with REL attached) VERB
```

```
nn
npp
art-301
identifier
774
dn
chr-000
wo7
dn
chr-001
jiyu'i
dcs2
art-303
Stem
chr-001
jiyu'i
dn
eng-000
ship
dn
eng-000
boat
dn
eng-000
airplane
nn
npp
art-301
identifier
775
dn
chr-001
jiyu junidisdi
dcs2
art-303
Stem
chr-001
jiyu junidisdi
dn
eng-000
airport
```

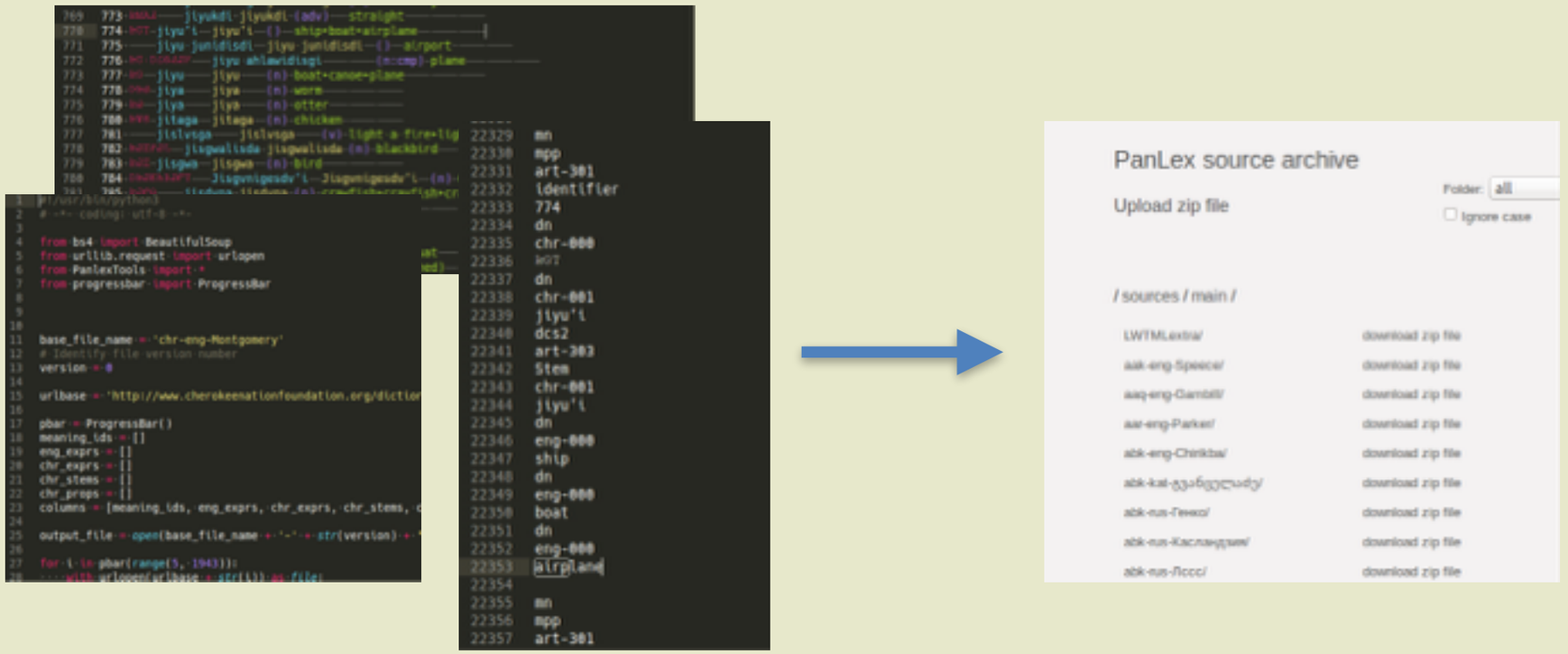
Case 1. Ben Yang: Cherokee

Submission to PanLex

- Send serialized data file to database



- Upload all data to PanLex source archive



Case 1. Ben Yang: Cherokee

Final result

PanLex 4.0

translation — see

from — expression

language		expression	
eng-000	English	541362	ship

into — expression

language — 114 — chr-000 — ᏣᎳᎩ ᎯᎠᏂᏱᎭᏘ

good	expression		see
7	22279204	ᏂᏩᏆ	

translation — translation

count

see

Case 1. Ben Yang: Cherokee

GV!

ᏀᏅᎯᏙᏣᏟᎢ ᏝᎬᏴᏂᎵᏍ!

Case 2. Gary Krug: Spanish dialects

- Examine formatting at all levels
  - Low-level format (html, xml, txt, etc.)

    acelerado = ***frenético***
- Structure of entries (what is included)
  - Extract part of speech and dialect

    sust. y adj. de abolengo (*bogotano*) => de abolengo [**sust,adj**]
- Change numbered markers and normalize delimiters

  1) con protuberancias 2) grumoso => con protuberancias, grumoso

Case 2. Gary Krug: Spanish dialects

- Solutions
  - Parenthesize extra words
    lleno de algo => **lleno** (*de algo*)
    to be **married**
  - Parenthesize helper words like "to" or "to be".
    estar feliz => (*estar*) **feliz**
  - Extract other characteristics (inchoative, causative)
    become **thirsty**
    make **rigid**

Voestes digibaagkoeh
(South Sámi – Norwegian Bokmål)

Case 3. Alex DelPriore: South Sámi

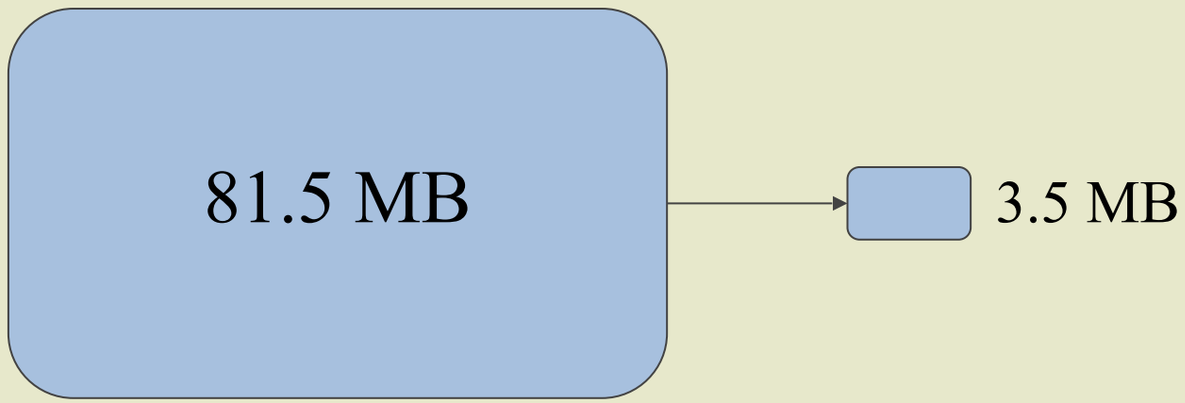
Lemma extraction

```
1255043  </small></small></span><span>Analyse: <small><i><c c="dimgray">indik. pret. 2p. dl.</c></i><c c="black">. ————</c></small>
         </span>
1255044  </indentIt></def>
1255045  ............
1255046  <head><k>vuejiehtidh</k>
1255047  </head><small><i>verb</i></small>
1255048  <small><small> Klasse <i>ulikest.</i></small></small>
1255049  <span>  1. <small><c c="dimgray"></c></small><span><bf>å </bf>jage</span> <small><small>
1255050  </small></small>  2. <small><c c="dimgray"></c></small><span><bf>å </bf>drive en flokk</span> <small><small>
1255051  </small></small></span><span>Analyser: <small><i><c c="dimgray">inf.</c></i><c c="black">. el. </c><i><c c="dimgray">indik.
         pret. 2p. pl.</c></i><c c="black">. ——</c></small></span>
1255052
1255053
1255054  <span><ln><indentIt>Nøkkelformer:
1255055  ·····<small><i><c c="dimgray">inf. ····</c></i></small><small><kref>vuejiehtidh</kref> </small>
1255056  ·····<small><i><c c="dimgray">indik. pres. 1p. sg. ····</c></i></small><small>(<kref>daenbiejjien</kref> <kref>manne</kref
         >) <kref>vuejehtem</kref> </small>
1255057  ·····<small><i><c c="dimgray">indik. pres. 3p. sg. ····</c></i></small><small>(<kref>daenbiejjien</kref> <kref>dïhte</kref
         >) <kref>vuejehte</kref> </small>
1255058  ·····<small><i><c c="dimgray">indik. pres. 3p. pl. ····</c></i></small><small>(<kref>daenbiejjien</kref> <kref>dah</kref>
         ) <kref>vuejiehtieh</kref> </small>
1255059  ·····<small><i><c c="dimgray">indik. pret. 1p. sg. ····</c></i></small><small>(<kref>jååktan</kref> <kref>manne</kref>) <
         kref>vuejiehtim</kref> </small>
1255060  ·····<small><i><c c="dimgray">neg. ····</c></i></small><small>(<kref>ij</kref>) <kref>vuejehth</kref> </small>
1255061  ·····<small><i><c c="dimgray">perf. ····</c></i></small><small>(<kref>lea</kref>) <kref>vuejiehtamme</kref> </small>
1255062  ·····<small><i><c c="dimgray">ger. ····</c></i></small><small>(<kref>lea</kref>) <kref>vuejehteminie</kref> </small>
1255063  ·····<small><i><c c="dimgray">verbgen. ····</c></i></small><small><kref>vuejehten</kref> </small>
1255064  </indentIt></ln></span><small><small>
1255065  </small></small><indentIt>Eksempler: ···</indentIt>
1255066  ····<small>vuejiehtidh bïrhketjem</small>
1255067  ····<i><small><c c="dimgray">drive den lille flokken</c></small></i>
1255068  ····<small>Datne hov mannem vuejehth!</small>
1255069  ····<i><small><c c="dimgray">Du driver jo meg!</c></small></i>
1255070
1255071  ............
1255072  <head><k>vuejiehtidie</k>
1255073  </head><small><i>verb</i></small><small> → </small><kref>vuejiehtidh</kref>
```

Case 3. Alex DelPriore: South Sámi

Post-extraction

```
10143  <html><head><k>vuejemelohkehtæjja</k></head><body><small><i>subst.</i></small><span> <small><c c="dimgray"></c></small><span>
       kjørelærer</span> <small><small></small></small></span><span>Analyse: <small><i><c c="dimgray">sg. nom.</c></i><c c="black">
       </c></small></span></body></html>
10144  <html><head><k>vuejiehtidh</k></head><body><small><i>verb</i></small><small><small> Klasse <i>ulikest.</i></small></small><
       span>   1. <small><c c="dimgray"></c></small><span><bf>å </bf>jage</span> <small><small></small></small>   2. <small><c c="
       dimgray"></c></small><span><bf>å </bf>drive en flokk</span> <small><small></small></small></span><span>Analyser: <small><i><c
       c="dimgray">inf.</c></i><c c="black"> el. </c><i><c c="dimgray">indik. pret. 2p. pl.</c></i><c c="black"> </c></small></span><
       small><small></small></small><indentit>Eksempler:    </indentit><small>vuejiehtidh birhketjem</small><i><small><c c="dimgray">
       drive den lille flokken</c></small></i><small>Datne hov mannem vuejehth!</small><i><small><c c="dimgray">Du driver jo meg!</c>
       </small></i></body></html>
10145  <html><head><k>vuejiehtæjja</k></head><body><small><i>subst.</i></small><span> <small><c c="dimgray"></c></small><span>en som
       driver flokken</span> <small><small></small></small></span><span>Analyse: <small><i><c c="dimgray">sg. nom.</c></i><c c="black
       "> </c></small></span></body></html>
```



Case 3. Alex DelPriore: South Sámi

Tabularization

```
<html>
<head><k>vuejiehtidh</k></head>
<body>
<small><i>verb</i></small>
<small><small> Klasse <i>ulikest.</i></small></small>
<span>  1. <small><c c="dimgray"></c></small><span><bf>å </bf>jage</span>
<small><small></small></small>
  2. <small><c c="dimgray"></c></small>
  <span><bf>å </bf>drive en flokk</span>
<small><small></small></small>
</span>
<span>Analyser: <small><i><c c="dimgray">inf.</c></i><c c="black"> el. </c><i><c c="dimgray">
indik. pret. 2p. pl.</c></i><c c="black"> </c></small></span>
<small><small></small></small>
<indentit>Eksempler:    </indentit>
<small>vuejiehtidh bïrhketjem</small>
<i><small><c c="dimgray">drive den lille flokken</c></small></i>
<small>Datne hov mannem vuejehth!</small>
<i><small><c c="dimgray">Du driver jo meg!</c></small></i>
</body>
</html>
```

Case 3. Alex DelPriore: South Sámi

Tabularization

TSV

expression (sma)	POS	grammar	class	expression (nob)
vuejedh	verb	inf.	I-å	kjøre
vuejeme	subst.	sg.▸nom.		kjøring
vuejeme	subst.	sg.▸nom.		tverrstang i gamme
vuejemeleahpa	subst.	sg.▸nom.		sertifikat▸førerkort
vuejemelohkehtæjja	subst.	sg.▸nom.		kjørelærer
vuejiehtidh	verb	inf.	ulikest-å	jage
vuejiehtidh	verb	inf.	ulikest-ä	drive en flokk
vuejiehtidh birhketjem				drive den lille flokken
Datne hov mannem vuejehth				Du driver jo meg
vuejiehtæjja	subst.	sg.▸nom.		en som driver flokken
vuejije	subst.	sg.▸nom.		sjåfør

Case 3. Alex DelPriore: South Sámi

Tabularization

TSV

expression (sma) **POS** **grammar** **class** **expression (nob)**

```
vuejedh-verb—inf.—I-(å) kjøre«dcs2:art-303»PartOfSpeechProperty«dcs:art-303»Verbal
vuejeme-subst.—sg.›nom.———kjøring
vuejeme-subst.—sg.›nom.———tverrstang i gamme
vuejemeleahpa-subst.—sg.›nom.———sertifikat›førerkort
vuejemelohkehtæjja—subst.—sg.›nom.———kjørelærer
vuejiehtidh-verb—inf.—ulikest-(å) jage«dcs2:art-303»PartOfSpeechProperty«dcs:art-303»Verbal
vuejiehtidh-verb—inf.—ulikest-(å) drive en
flokk«dcs2:art-303»PartOfSpeechProperty«dcs:art-303»Verbal
vuejiehtidh bïrhketjem————drive den lille flokken
Datne hov mannem vuejehth————Du driver jo meg
vuejiehtæjja—subst.—sg.›nom.———en som driver flokken
vuejije-subst.—sg.›nom.———sjåfør
```

Case 3. Alex DelPriore: South Sámi

Serialization

subst.

```
242514  mn
242515  dn
242516  sma-000
242517  vuejemelohkehtæjja
242518  dcs2
242519  art-303
242520  PartOfSpeechProperty
242521  art-303
242522  Noun
242523  dn
242524  nob-000
242525  kjørelærer
```

```
242527  mn
242528  dn
242529  sma-000
242530  vuejiehtidh
242531  dcs2
242532  art-303
242533  PartOfSpeechProperty
242534  art-303
242535  Verbal
242536  dpp
242537  art-303
242538  Class
242539  ulikest
242540  df
242541  nob-000
242542  (å)·jage
242543  dn
242544  nob-000
242545  jage
242546  dcs2
242547  art-303
242548  PartOfSpeechProperty
242549  art-303
242550  Verbal
```

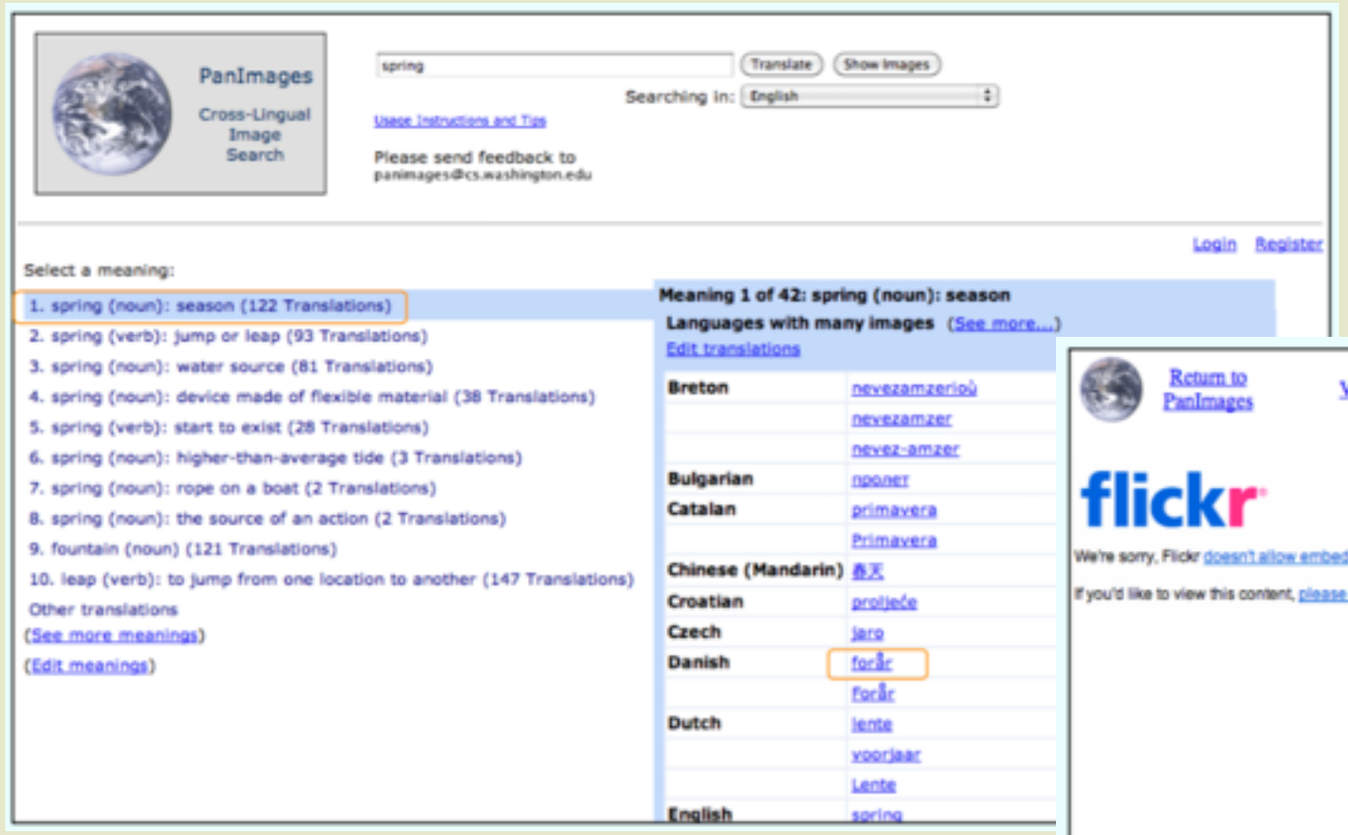
verb

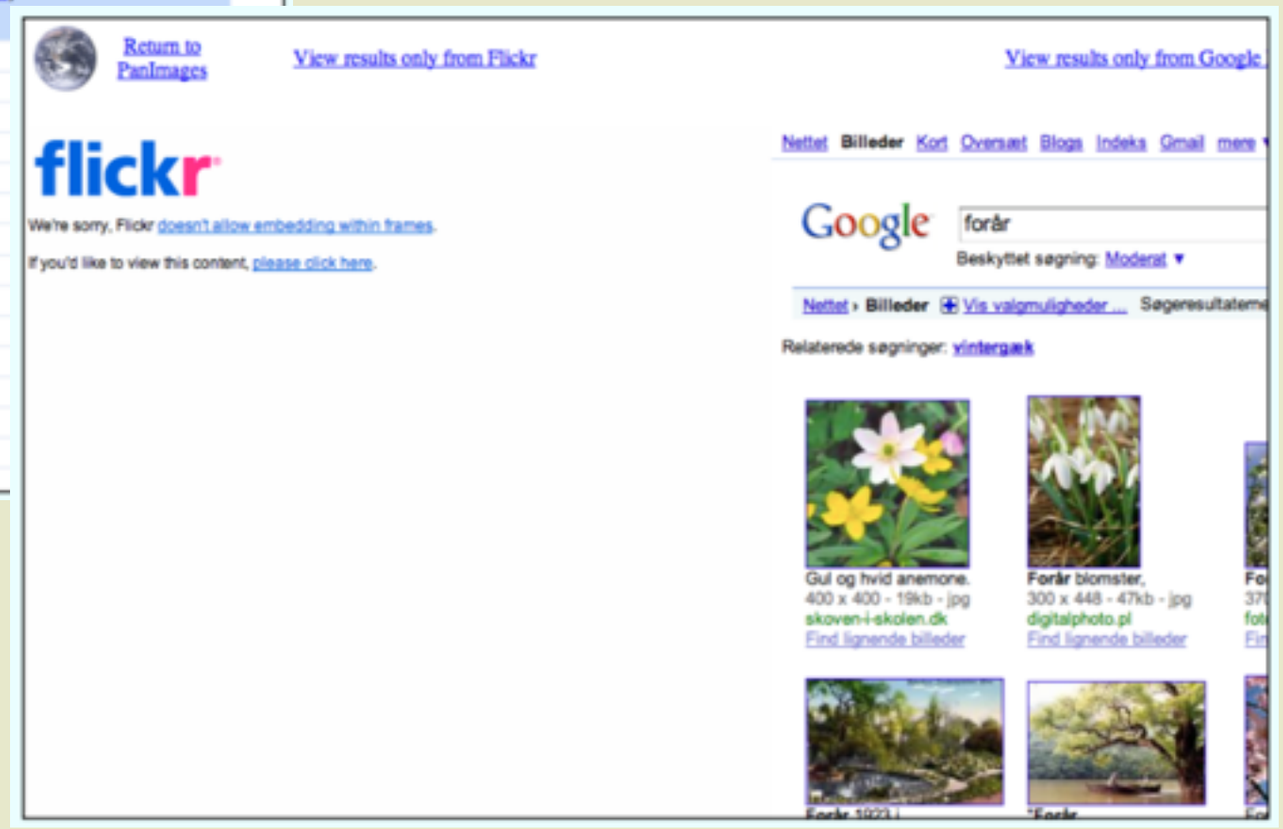
```
242552  mn
242553  dn
242554  sma-000
242555  vuejiehtidh
242556  dcs2
242557  art-303
242558  PartOfSpeechProperty
242559  art-303
242560  Verbal
242561  dpp
242562  art-303
242563  Class
242564  ulikest
242565  df
242566  nob-000
242567  (å)·drive·en·flokk
242568  df
242569  nob-000
242570  drive·en·flokk
```

reclassified

Results

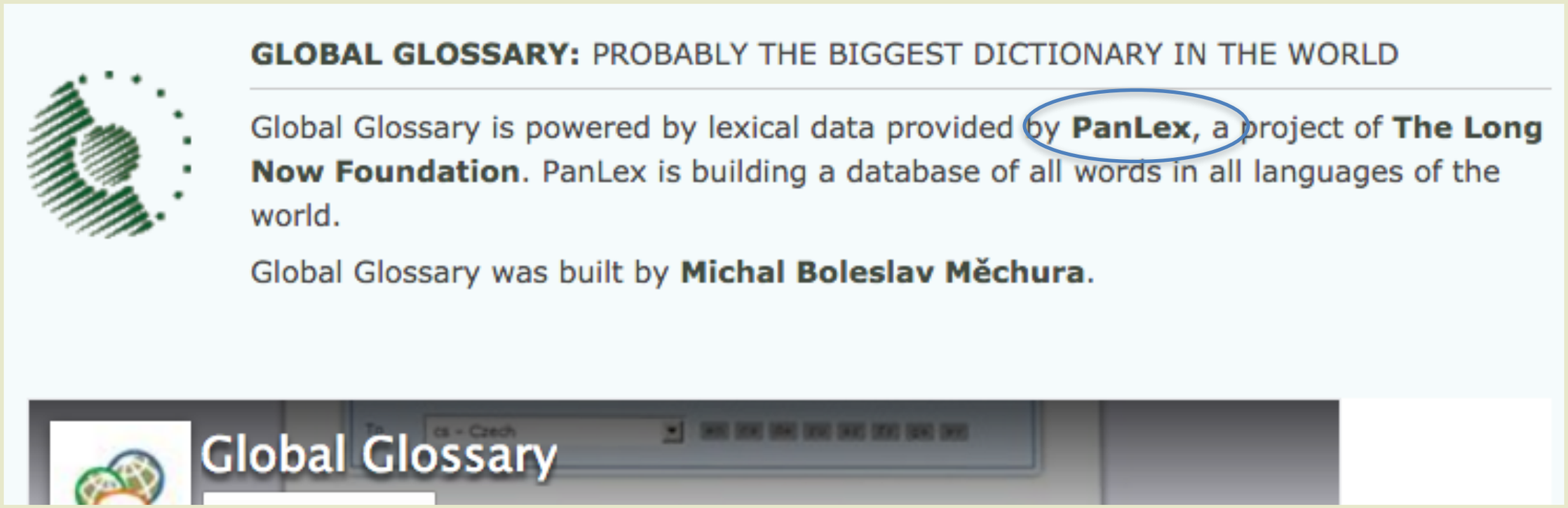
- Demonstrated use cases
  - PanImages





Results

- Demonstrated use cases
  - Global Glossary



Results

- Researched and published:
  - Westphal *et al.*, "Countering language attrition with PanLex and the Web of Data" (02015)
  - Kamholz *et al.*, "PanLex: Building a Resource for Panlingual Lexical Translation" (02014)
  - Gola, "An analysis of translation divergence patterns using PanLex translation pairs" (02012)
  - Mausam *et al.*, "Panlingual Lexical Translation via Probabilistic Inference" (02010)
  - Pool, "Panlingual Globalization" (02010)
  - Everitt *et al.*, "Evaluating Lemmatic Communication" (02010)
  - Baldwin *et al.*, "PanLex and LEXTRACT: Translating all Words of all Languages of the World" (02010)
  - Soderland et al., "Lemmatic Machine Translation" (02009)
  - Christensen *et al.*, "A Rose is a Roos is a Ruusu: Querying Translations for Web Image Search" (02009)
  - Colowick, "Multilingual Search with PanImages" (02008)
  - Etzioni *et al.*, "Lexical Translation with Application to Image Search on the Web" (02007)

Results

- Acquired 6K sources (http://panlex.org/tech/plrefs.shtml)
- Consulted 2K sources
- Built a database with
  - 10K varieties of 6K languages
  - 22M expressions
  - 1.2B distance-1 translations
  - 33B distance-2 translations

Access

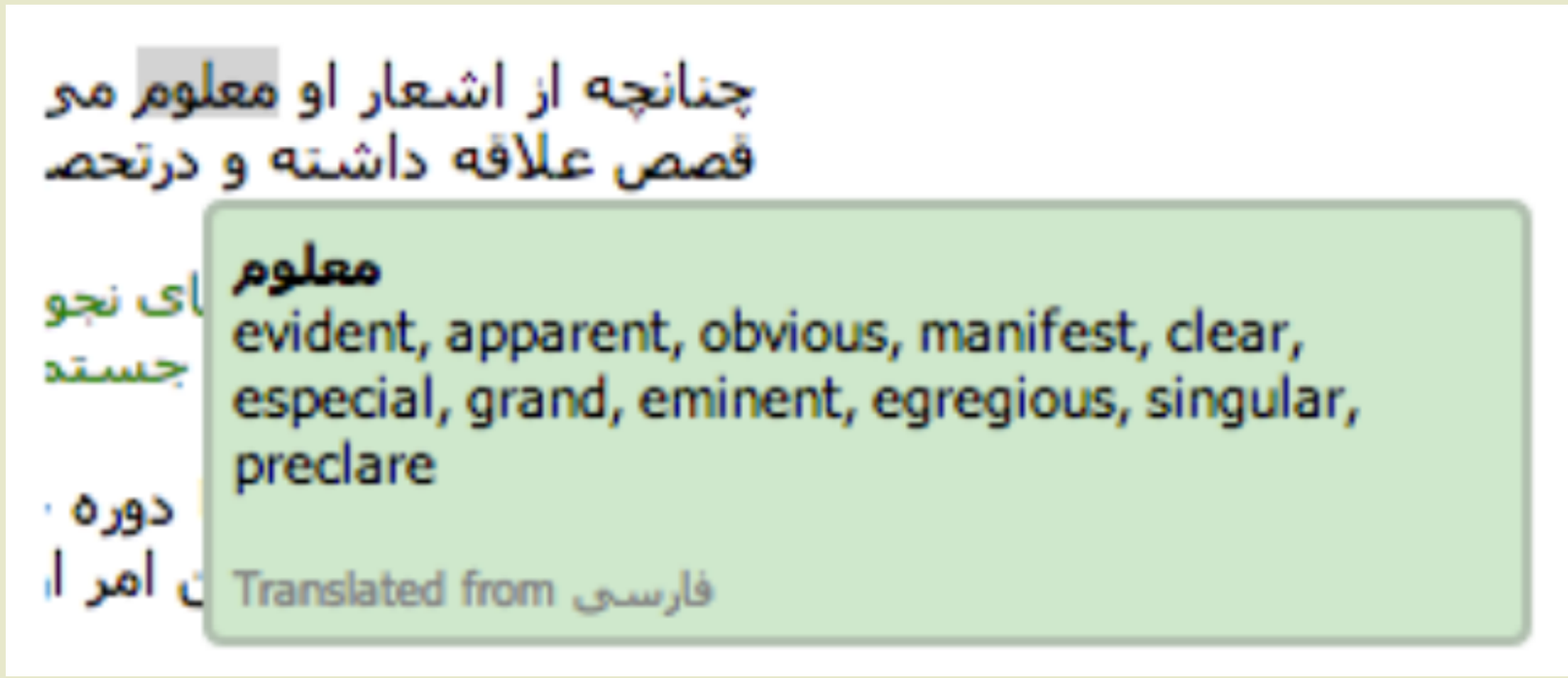
- TeraDict

Live demo

http://panlex.org/teradict/?lg=eng

Access

- PanLex Translator (Chrome extension)



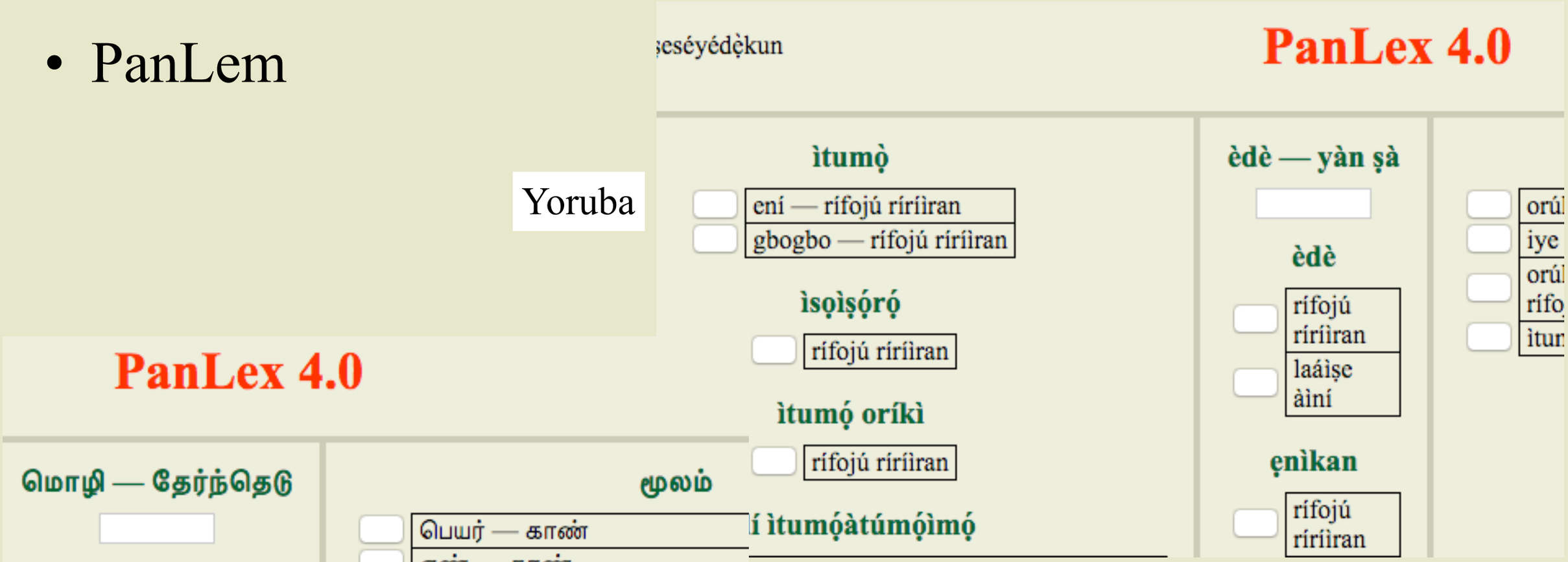
Access

- PanLinx

Access

- PanLem

Yoruba



Tamil

Access

- NLTK (Natural Language Toolkit)
  - Swadesh-Yakhontov 110 List: 2,000 languages
  - Swadesh 207 List: 800 languages

41. *Open Multilingual Wordnet* [download | source]
    id: omw; size: 25057024; author: Francis Bond; copyright: Please consult the copyright statements of the indivi
    consult the LICENSE files included with the individual Wordnets. Note that all permit redistribution.;

42. *Opinion Lexicon* [download | source]
    id: opinion_lexicon; size: 24947; author: Bing Liu; copyright: Copyright (C) 2011 Bing Liu; license: Creativ
    International;

43. *PanLex Swadesh Corpora* [download | source]
    id: panlex_swadesh; size: 2699578; author: Jonathan Pool (editor); copyright: ; license: CC0 1.0 Universal;

44. *Paradigm Corpus* [download | source]
    id: paradigms; size: 24902; author: Cathy Bow, University of Melbourne; copyright: ; license: Distributed wit

Access

- PanLex API
  - Available via the HTTP endpoint api.panlex.org
  - JSON queries and responses
  - Public (limit of 2 queries/sec)
  - Exposes most database objects (expressions, sources, meanings, etc.)
  - Provides distance-1 and distance-2 translations with quality scores

Access

- PanLex API
  - Example query

```
$ curl http://api.panlex.org/ex -d
'{ "uid": "zul-000", "truid": "gle-000",
"trtt": "rialtas", "trdistance": 2,
"include": "trq", "sort": "trq desc",
"limit": 5 }'
```

Access

- PanLex API
  - Example response


```
{
  "resultType": "ex",
  "result": [
    {
      "ex": 1297265,
      "lv": 832,
      "tt": "uhulumeni",
      "td": "uhulumeni",
      "trex": 878773,
      "trq": 574
    }
  ],
  "resultNum": 1,
  "resultMax": 2000
}
```

Access

- Dumps
- PanLex RDF
- PanLex Tattoo Generator

Help us

- Helping us document the long tail of neglected languages
  - Finding digital and print dictionaries and word lists
  - Helping with crowdsourcing planning
- Contributing ready-to-ingest data

Help us

- Designing apps and interfaces
  - Localization tools
  - Mobile apps
  - Graph visualizations
  - Adding links to other (linguistic, geographic, etc.) data
- Investigating translation inference algorithms

Discussion

- Use cases?
- Suggestions?
- Questions?
- Want to help?