

<http://panlex.org>

ἘΦΤΑΣΗΕ Localization:
Translating Every Word
in Every Language

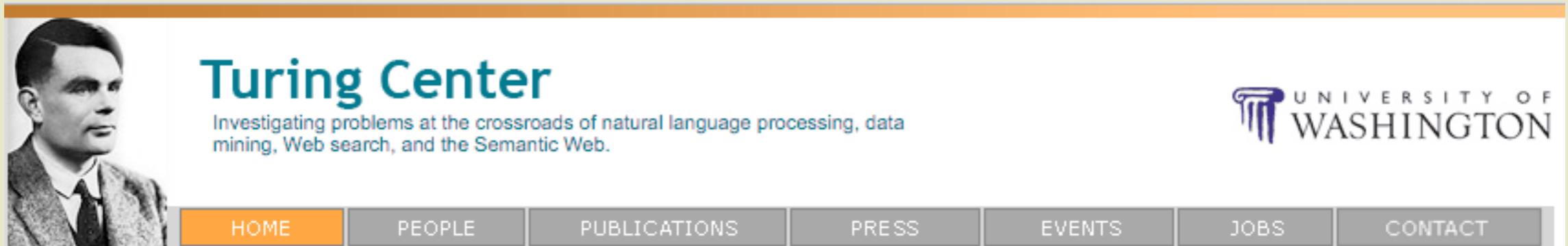
Outline

- 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:



Who we are

- From 02012 to now at:



Seminars About Long-term Thinking
Hosted by Stewart Brand. This series is building a coherent, compelling body of ideas about long-term thinking.

The Interval
A compelling venue for conversation that invites visitors to spend time in a place that itself encourages long-term thinking.

10,000 Year Clock
The 10,000 Year Clock Project was conceived by Danny Hillis as a monument to long-term thinking.

The Rosetta Project
The Rosetta Project is now the largest collection of linguistic data on the Net.

PanLex
PanLex is creating a collection of all the words of all the world's languages...

Long Server
The overarching program for our digital continuity software projects.

Revive & Restore
Genetic rescue for endangered and extinct species.

Long Bets
Make predictions or bets about future events of importance.

Nevada
We have purchased desert mountain land adjoining Great Basin National Park in eastern Nevada.

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

Mausam

Stephen Soderland

Kobi Reiter

Michael Skinner

Marcus Sammer

Timothy Baldwin

Katherine Everitt

Christopher Lim

Janara Christensen

Daniel S. Weld

Jeff Bilmes

Katrin Kirchhoff

Bo Qin

- 02010–now (* = here at this talk):

Jonathan Pool*

David Kamholz*

Susan Colowick

Alex DelPriore*

Gary Krug*

Benjamin Yang*

Julie Anderson*

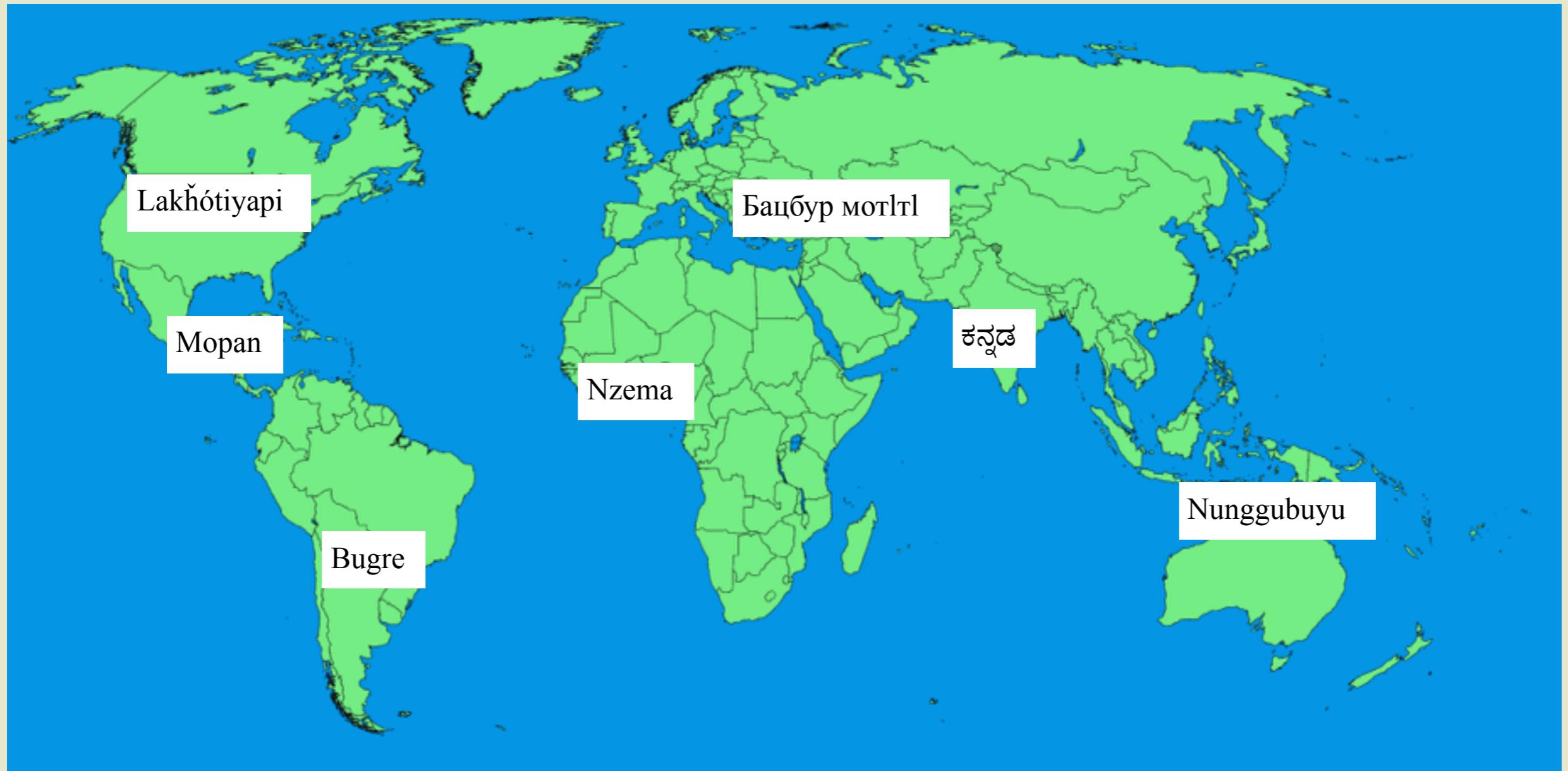
Emily Bender

Steven Bird

- And interns, volunteers, advisors

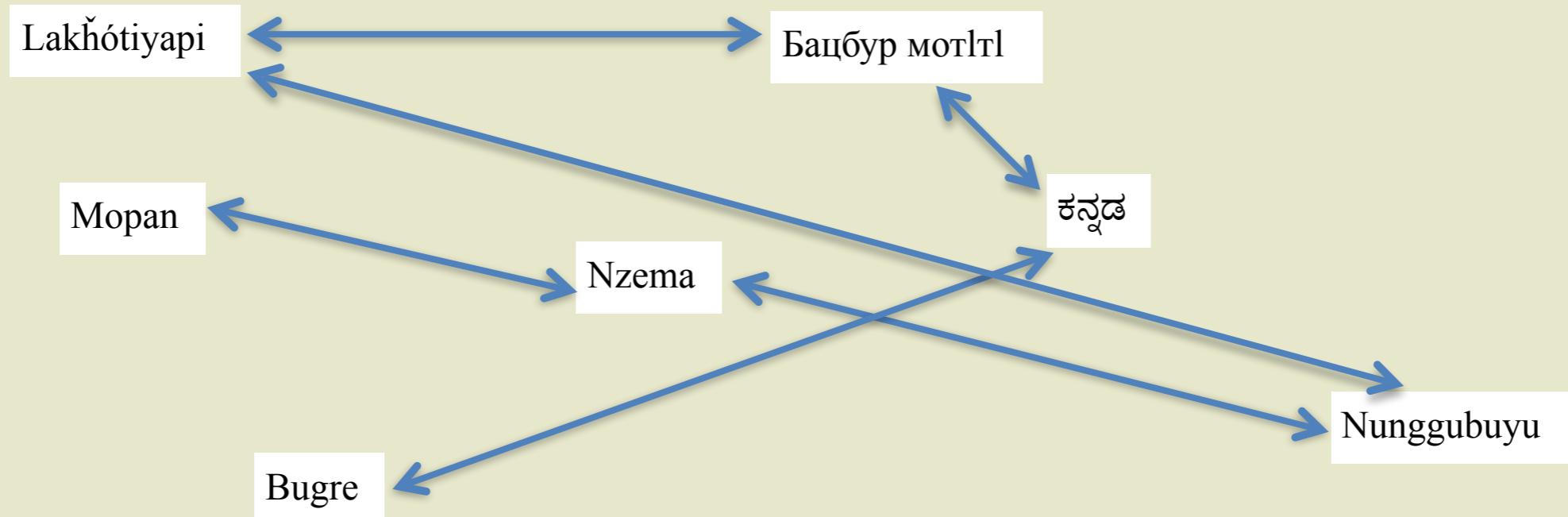
Vision

- Every language a world language.



Vision

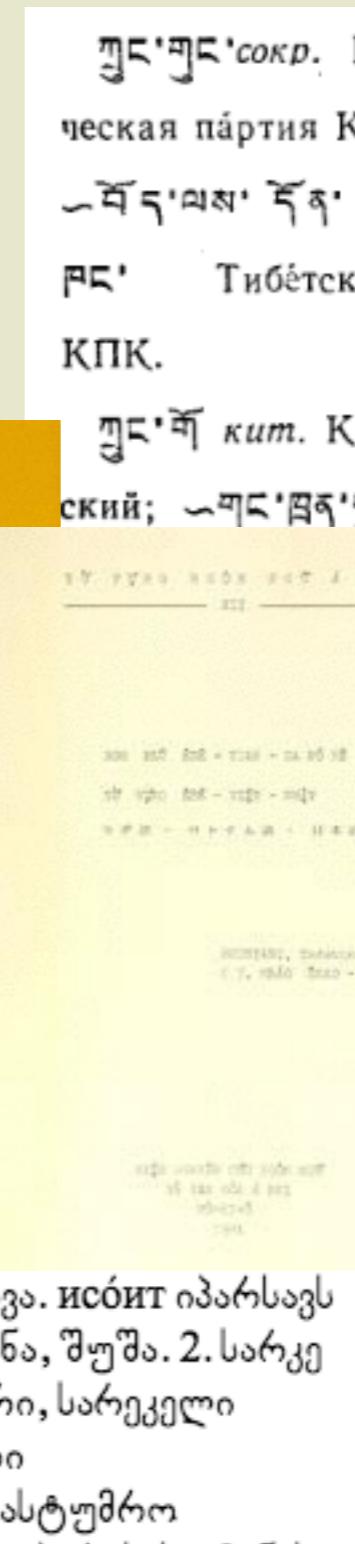
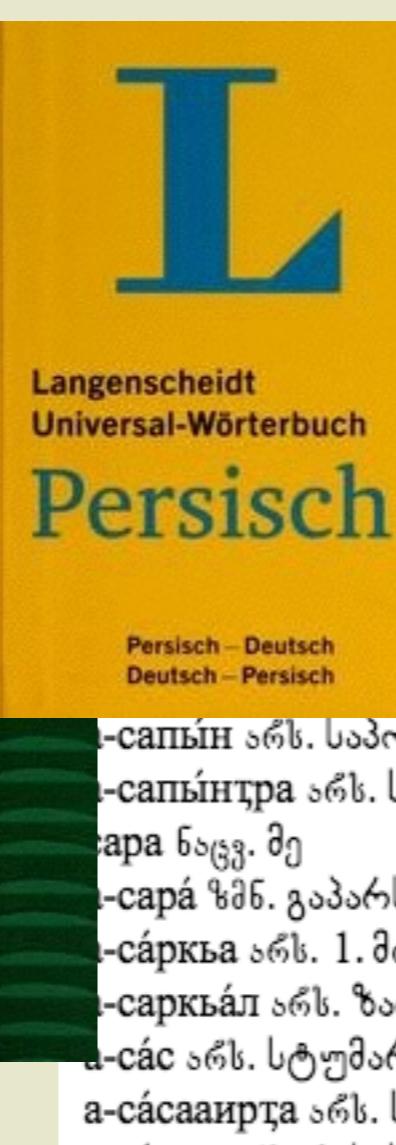
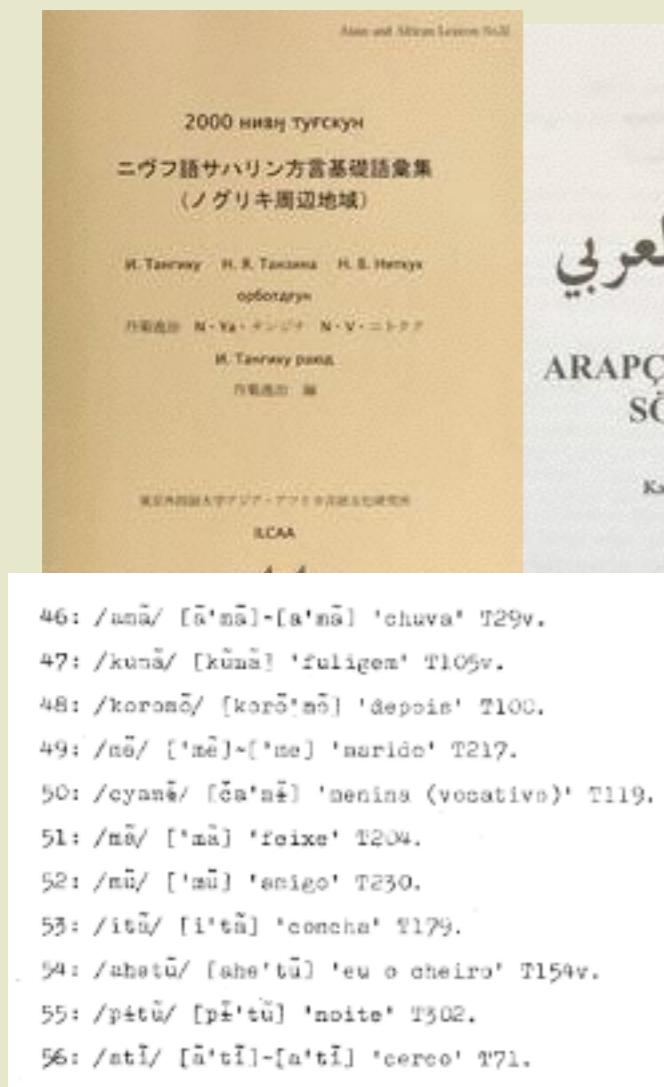
- Any language translatable into any other language.



$\approx 7K$ languages $\Rightarrow 50M$ pairs

Strategy

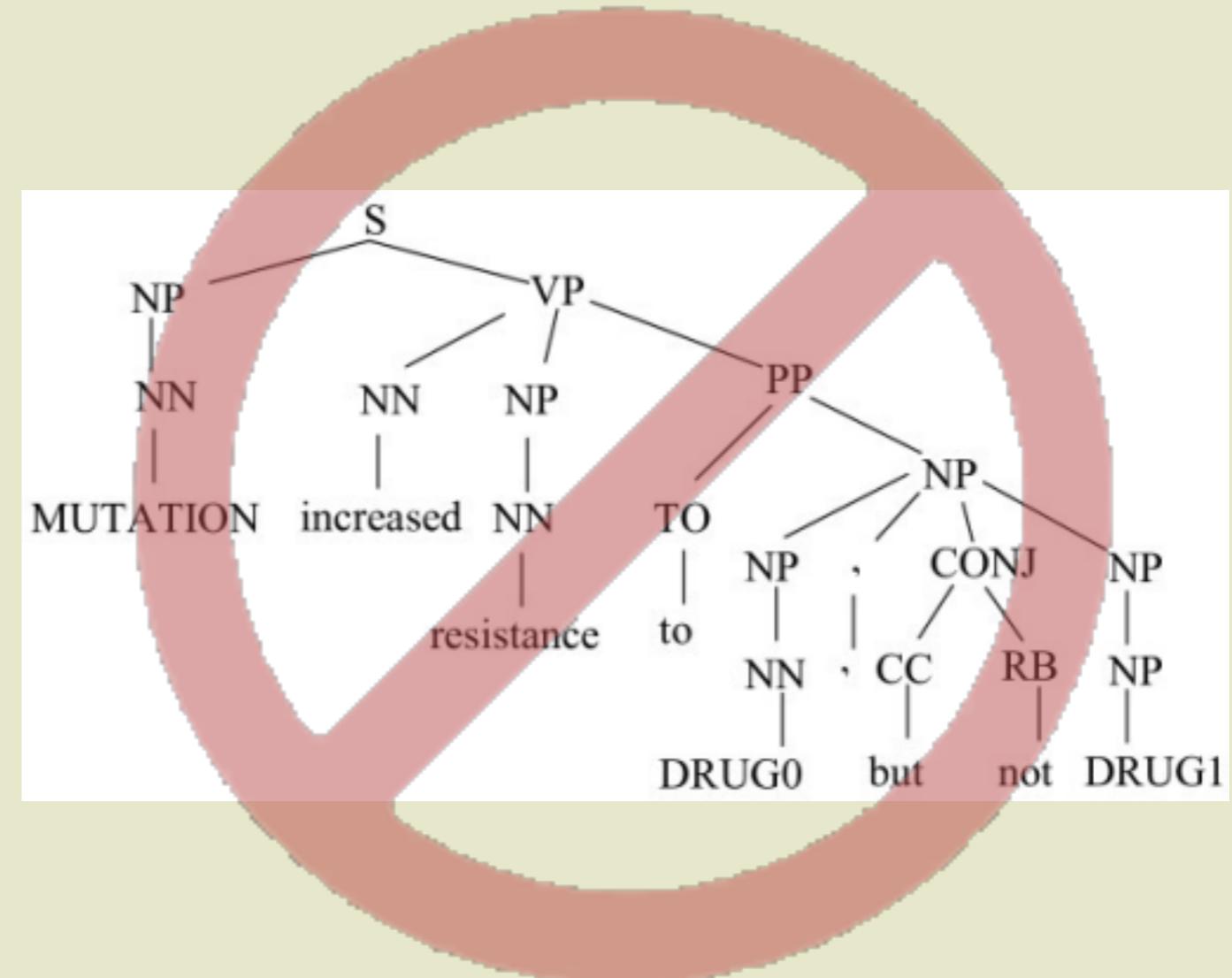
1. Acquire existing data first
2. Preserve and acknowledge each fact's provenance



Strategy

3. Start with lexicon, defer grammar

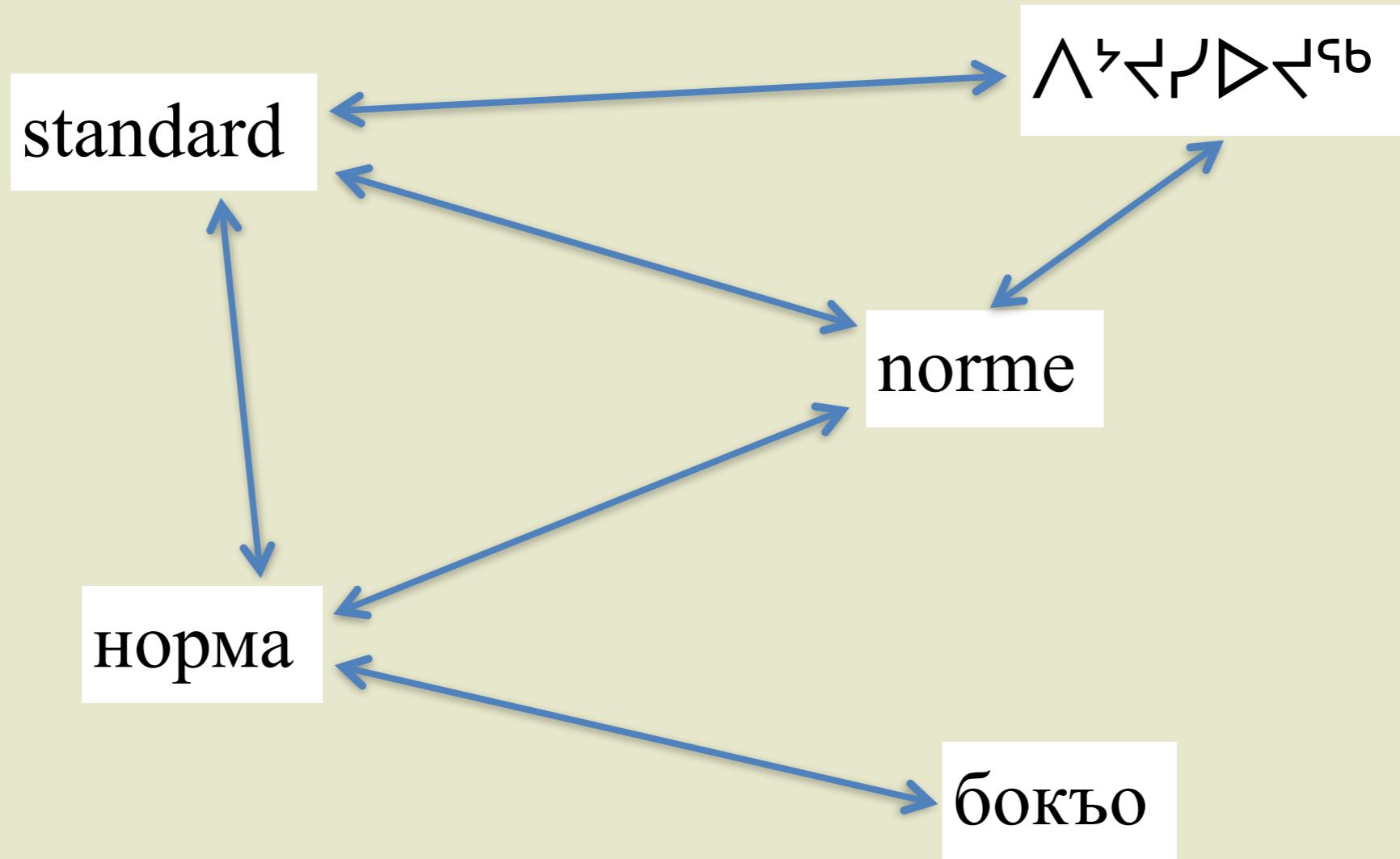
17596085 qaq
17605846 qaritte=ma
17603016 qaro
17369810 qaru
17109503 qarun-kó
17577214 qarunko
17577239 qarunte
17585713 qaw
17222182 qaw-
17093499 qaw-xó
17222324 qawai
17095202 qawho
17605754 qawtitto
17221018 qaww-
17093643 qawxó
17342725 qawčitto
17606812 qayanko
17575892 qayanqayo
17607397 qayangayo



Strategy

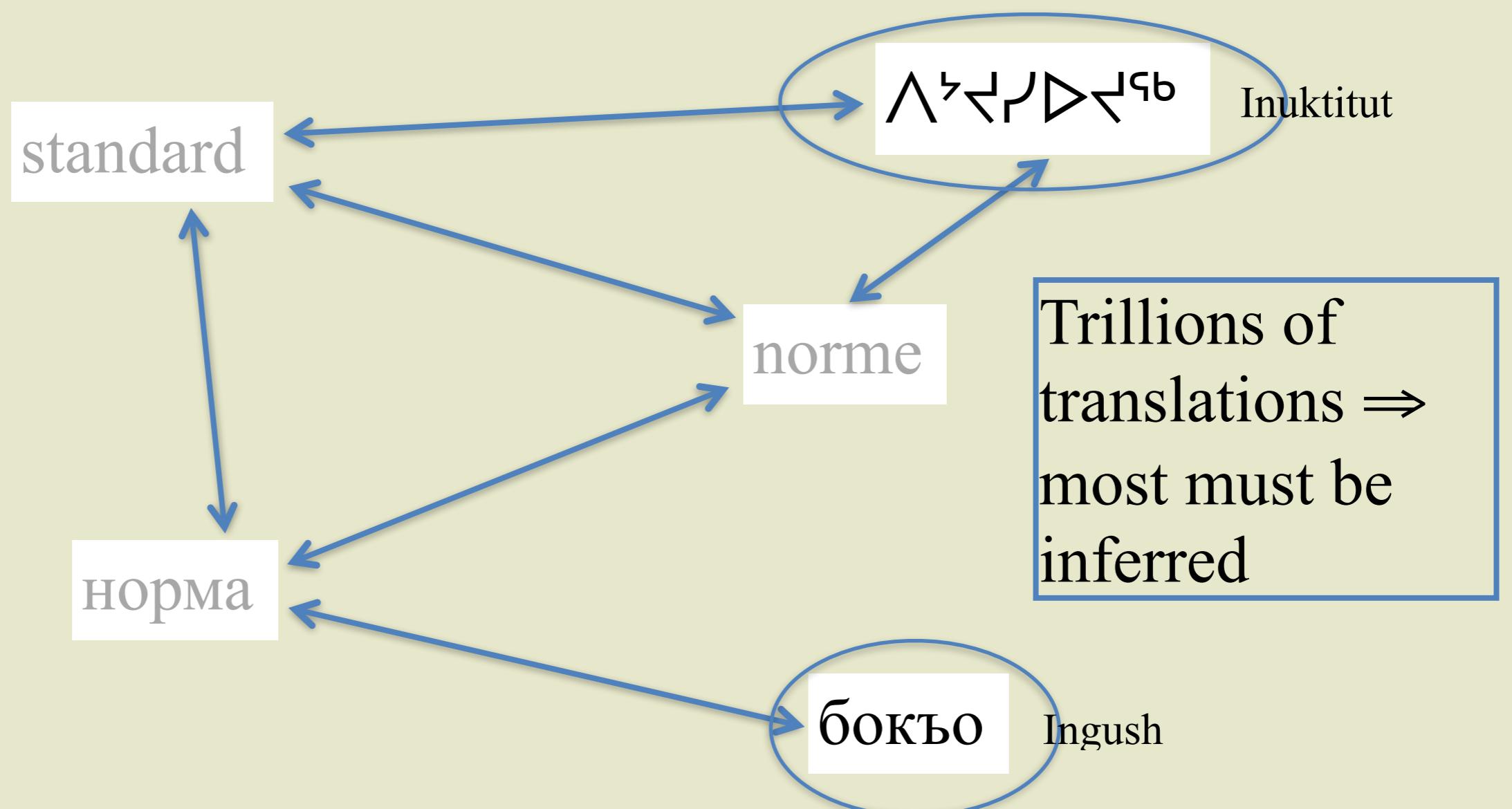
4. Translate:

- Lexemes
- Symmetrically
- Any-to-any



Strategy

5. Infer missing translations:



Concepts

Languages have varieties.

| | |
|---------|-----------------|
| azj-000 | azərbaycanca |
| azj-001 | Азәрбајҹан дили |
| azj-002 | Төрөкөмөсө |
| azj-003 | терекеме |
| azj-004 | آذربایجانی |

Language varieties have *expressions*.

| | |
|----------|----------------------|
| 18433157 | ишдә |
| 18434129 | ишдән бојун гачырмаг |
| 18434128 | ишдән гачмаг |
| 18438369 | иши чох олан |
| 18434163 | ишелдэлилмиш |
| 18434157 | ишелдэлилмәк |
| 18453099 | ишелмә |

Concepts

Expressions have *meanings*.

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

Each meaning has a *source*.

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

source — 3081 — pum-eng-npi:Rai
meaning — 19280118

| denotation | | | | | more |
|------------|----------|------------|----------|------------|--------------------------|
| code | language | expression | | | |
| 48076120 | npi-000 | नेपाली | 2604048 | छुचुन्द्रो | <input type="checkbox"/> |
| 48076119 | pum-000 | Pumā | 17568478 | cʌktoŋ | <input type="checkbox"/> |

source — 372 — npi-epo:Mandahar
meaning — 4519169

| denotation | | | | | more |
|------------|----------|------------|---------|------------|--------------------------|
| code | language | expression | | | |
| 13528666 | epo-000 | Esperanto | 650223 | talpo | <input type="checkbox"/> |
| 13528665 | npi-000 | नेपाली | 2604048 | छुचुन्द्रो | <input type="checkbox"/> |

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

C - C

ɔbacá *n.* bouton de gale.

ɔbála *n.* chaleur, fièvre.

ɔbalasaná *n.* sorte de terme qui apparaît en nov. (à partir de 16h.).

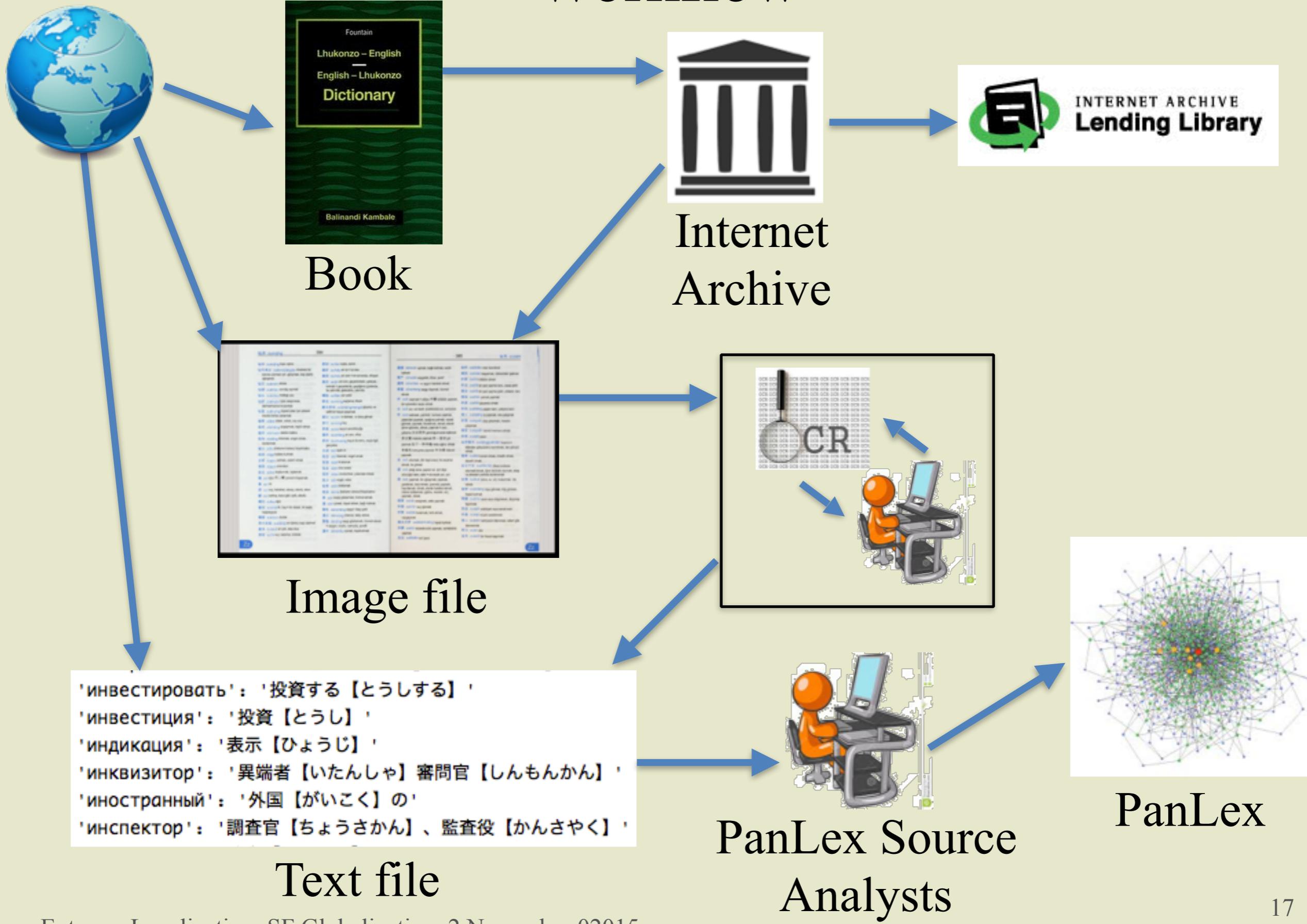
ɔbéle *n.* trou de fourmis.

ɔbéle *n.* trou de sortie des termites.

ɔbɔbe *n.* charme.

| | |
|------------------------------|---|
| 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 nɔmaándé - 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 | |

Workflow



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'.



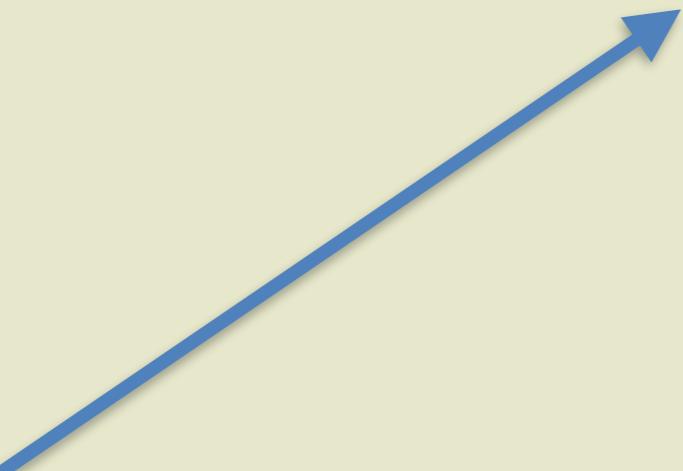
| | | |
|--------|------|-----------------|
| balima | noun | heaven▶sky camp |
| balima | noun | place far away |

Workflow

- Serialization:

| | | |
|--------|------|----------|
| balima | noun | heaven |
| balima | noun | sky camp |

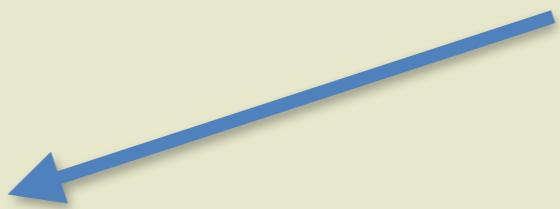
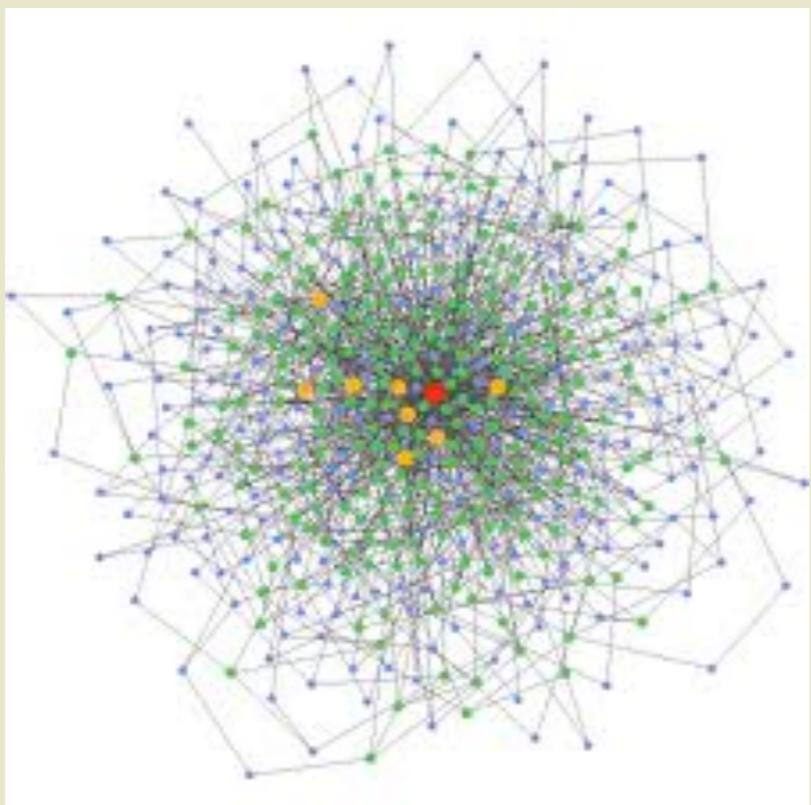
place far away



```
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

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 | Издательство “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:ТобылъТ҃ | 7 | Къарча-Черкес Респуб | | abq-002[1]; abq-003[1] |
| | abq:ТобылъЩ҃ | 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

PanLex source archive

Upload zip file

Folder: all

Ignore case

Search

/sources /main /

| | |
|---------------------|-----------------------------------|
| LWTMLextra/ | download zip file |
| azk-eng-Specel/ | download zip file |
| aqq-eng-Gambili/ | download zip file |
| azx-eng-Parker/ | download zip file |
| abk-eng-Chirkba/ | download zip file |
| abk-kaz-бэлгүүчэдг/ | download zip file |
| abk-nis-Генко/ | download zip file |
| abk-nis-Каслаандэм/ | download zip file |
| abk-nis-Лосо/ | download zip file |
| abk-nis-Mn/ | download zip file |
| abk-nis-Cn/ | download zip file |
| abk-nis-TA/ | download zip file |
| abk-nis-TrV/ | download zip file |
| abk-nis-Хөхөн/ | download zip file |
| abk-nis-ZI/ | download zip file |
| abk-nis-эл-Русские/ | download zip file |

CVY D&P.0 SJeLi
Cherokee Nation Foundation

Home Cherokee Pronouns How to use

Archive for March, 2011

< Previous Entries
-T-1 (1)
Today, March 29th, 2011

English entry: LDC Locative
Cherokee: -T-1
Stem: T1

| Syllables | Phonetics | English |
|-----------|-----------------------|---|
| here | Jasgwoh! i dagvynill. | I'm going to hit you in the stomach. GAW'CT WEENP. |

Posted in grammar | No Comments »

JOSWILY DOWASPAKE dinadehlgwasgi analegalozi untagewathiv ()
Today, March 29th, 2011

English entry: homecoming parade

Case 1. Ben Yang: Cherokee

Data extraction

The screenshot shows a web page from the Cherokee Nation Foundation's dictionary. The header includes the logo of the Cherokee Nation and the text "GWY DBPÅ SJÅDU". The search bar contains "Search Type and Entry". Below the search bar, a red box highlights the entry for "uska". The entry details are as follows:

- English entry:** head; skull
- Cherokee:** OGAID uska
- Silm:** ska(n)

The entry is listed in the "Dictionary" section of the table. The table has three columns: "Dictionary", "Properties", and "English". The "Dictionary" column contains "uska". The "Properties" column contains "na uska agweda hagena.". The "English" column contains "Target practice on that bald head.".

At the bottom of the page, there is a note: "This entry was posted on Friday, March 20th, 2015 at 10:00 am and is filed under Books. You can follow any responses to this entry through the RSS 2.0 feed. You can leave a response, or trackback from your own site." There is also a link to "View original post".

The screenshot shows a terminal window with Python code. The code is used to extract data from the Cherokee dictionary website. The code uses BeautifulSoup to parse the HTML of the dictionary pages and extract specific entries. The code is as follows:

```
1 #!/usr/bin/python3
2 # -*- coding: utf-8 -*-
3
4 from bs4 import BeautifulSoup
5 from urllib.request import urlopen
6 from PanlexTools import *
7 from progressbar import ProgressBar
8
9
10
11 base_file_name = 'chr-eng-Montgomery'
12 # identify file version number
13 version = 8
14
15 urlbase = 'http://www.cherokeenationfoundation.org/dictionary/?p='
16
17 pbar = ProgressBar()
18 meaning_ids = []
19 eng_exps = []
20 chr_exps = []
21 chr_stems = []
22 chr_props = []
23 columns = [meaning_ids, eng_exps, chr_exps, chr_stems, chr_props]
24
25 output_file = open(base_file_name + '-' + str(version) + '.txt', 'w', encoding='utf-8')
26
27 for i in pbar(range(5, 1943)):
28     with urlopen(urlbase + str(i)) as file:
29         soup = BeautifulSoup(file.read())
30         meaning_ids.append(str(i))
31         eng_exps.append(soup.find('span', {'class': 'glo'}).text)
32         chr_exps.append(soup.find('span', {'class': 'def'}).text)
33         chr_stems.append(soup.find('span', {'class': 'stm'}).text)
34         chr_props.append(soup.find('span', {'class': 'partsp'}).text)
35
36 tabbed_output(columns, output_file)
37
38 output_file.close()
```

Case 1. Ben Yang: Cherokee

Tabularization

- Organize data

| | | | | |
|------------|-----|-------------------------------|------------------|--|
| meaning id | 555 | air, storm , tornado, wind | chr-000 (CWY) | Cherokee Stem |
| | | O ^o Z ^o | unole | hole (n) chr-001 (Tsalagi) Cherokee Part of speech |

- Fix regular errors
 - t → L
- 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

```
391 verb-suffix—art-303:Morpheme:art-303:Suffix—  
392 verb-intransitive—art-303:PartOfSpeechProperty:art-303:IntransitiveVerb—  
393 verb-transitive art-303:PartOfSpeechProperty:art-303:TransitiveVerb—  
394 verbs—art-303:PartOfSpeechProperty:art-303:Verbal—  
395 Verbs art-303:PartOfSpeechProperty:art-303:Verbal—
```

- Normalization
- Output

```
769 773-HMA:t—jiyukdi-jiyukdi-(adv)—straight—'  
770 774-HGT-jiyu'i-jiyu'i—()—ship+boat+airplane—|  
771 775—jiyu:junidisdi—jiyu:junidisdi—()—airport—  
772 776-HG-00842F—jiyu:ahlawidisgi—(n:cmp)-plane—  
773 777-HG—jiyu—jiyu—(n)-boat+canoe+plane  
774 778-Dna-jiya—jiya—(n)-worm—  
775 779-Hd—jiya—jiya—(n)-otter—  
776 780-HVz-jitaga—jitaga—(n)-chicken—  
777 781—jislvsqa—jislvsqa—(v)-light-a-fire+light-a-match—VERB—  
778 782-hazza:jisgwalsda-jisgwalsda-(n)-blackbird—  
779 783-haz-jisgwa—jisgwa—(n)-bird—  
780 784-Dmabik2mt—Jisgvnigesdv'i—Jisgvnigesdv'i—(n)-Gore-(place-of carp)—  
781 785-hora—jisdvna jisdvna (n)-crawfish+crayfish+crawdad—  
782 786-HMz-jlsdu:jlsdi-jlsdu+jlsdi-(n)-rabbit—  
783 787-HH—jimi—jimi—(n)-Jimmy—  
784 788—jllvsgl-jllvsgl-(n)-flower—  
785 789—degatsilugv (CL)—jihlusga—(v)-squat—VERB—  
786 790-HY—jigi—jigi—(v)-be-(with REL-attached)—VERB—
```



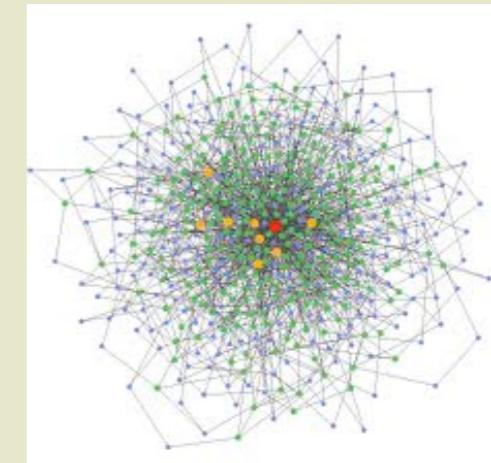
```
22329 nn  
22330 npp  
22331 art-303  
22332 identifier  
22333 774  
22334 dn  
22335 chr-008  
22336 ierT  
22337 dn  
22338 chr-001  
22339 jiyu'l  
22340 dcs2  
22341 art-303  
22342 Stem  
22343 chr-001  
22344 jiyu'l  
22345 dn  
22346 eng-008  
22347 ship  
22348 dn  
22349 eng-008  
22350 boat  
22351 dn  
22352 eng-008  
22353 airport  
22354 nn  
22355 npp  
22356 art-303  
22357 Identifier  
22358 775  
22359 dn  
22360 chr-001  
22361 jiyu:junidisdi  
22362 dcs2  
22363 art-303  
22365 Stem  
22366 chr-001  
22367 jiyu:junidisdi  
22368 dn  
22369 eng-008  
22370 airport  
22371
```

Case 1. Ben Yang: Cherokee

Submission to PanLex

- Send serialized data file to database

22329 nn
22330 npp
22331 art-381
22332 tIdentifier
22333 774
22334 dn
22335 chr-666
22336 regT
22337 dn
22338 chr-661
22339 jiyu't
22340 dcs2
22341 art-383
22342 Stem
22343 chr-661
22344 jiyu'i
22345 dn
22346 eng-666
22347 shlp
22348 dn
22349 eng-666
22350 boat
22351 dn
22352 eng-666
22353 airplane
22354 nn
22355 npp
22356 art-381



- Upload all data to PanLex source archive

```
773 773-MAN---jiyukdi-jiyukdi-(adv)---straight
774 774-HGT-jiyu'l-jiyu'-()---ship=boat+airplane
775 775---jiyu junidisdi---jiyu junidisdi---()---airport
776 776-HG-SCA2P---jiyu whamwidagi---(inc-ppl)-plane
777 777-HG-jilya---jiyu---(n)-boat+canoe+plane
778 778-DHD-jilya---jiyu---(n)-wore
779 779-ha-jilya---jiyu---(n)-otter
780 780-HHH-jiltega-jiltega---(n)-chicken
781 781-jislvsga---jislvsga---(vv-light-a fire+litg
782 782-HHHHL-jisgvnigeadu-jisgvnigeadu---(n)-blackbird
783 783-hd-jisgwa---jisgwa---(n)-bird
784 784-CHUNKER---Jisgvnigeadu'L-Jisgvnigeadu'L---(n)-i
785 785-WNG-WNG-----WNG-WNG
1 prusr/blackpython
2 # -*- coding: utf-8 -*-
3
4 from bs4 import BeautifulSoup
5 from urllib.request import urlopen
6 from PanlexTools import *
7 from progressbar import Progressbar
8
9
10 base_file_name = 'chr-eng-Montgomery'
11 # Identify file version number
12 version = 0
13
14 urlbase = 'http://www.cherokeenationfoundation.org/diction
15
16 pbar = Progressbar()
17 meaning_ids = []
18 eng_expressions = []
19 chr_expressions = []
20 chr_stems = []
21 chr_prepes = []
22 columns = [meaning_ids, eng_expressions, chr_expressions, chr_stems, c
23
24 output_file = open(base_file_name + '-' + str(version) + '-'
25
26 for i in pbar(range(5, 1943)):
27     with urlopen(urlbase + str(i)) as file:
```



| PanLex source archive | |
|--|---|
| <input type="checkbox"/> Folder: | <input checked="" type="checkbox"/> all |
| <input type="checkbox"/> Upload zip file | <input type="checkbox"/> Ignore case |
| <i>/sources/main/</i> | |
| LWTMlextra/ | download zip file |
| aak-eng-Speccer/ | download zip file |
| aaq-eng-Gambar/ | download zip file |
| aar-eng-Parker/ | download zip file |
| abk-eng-Chirkba/ | download zip file |
| abk-kaz-әзәрбайжан/ | download zip file |
| abk-nus-Генко/ | download zip file |
| abk-nus-Касландин/ | download zip file |
| abk-nus-Лисс/ | download zip file |

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 — GWY ᏃᎳh.Ꮡ.

| good | expression | see |
|------|------------|------|
| 7 | 22279204 | hrGT |

translation — translation

count

see

Case 1. Ben Yang: Cherokee

GV!

GoVGT LEBhP!

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**

Case 3. Alex DelPriore: South Sámi

Voestes digibaagkoeh
(South Sámi – Norwegian Bokmål)

Case 3. Alex DelPriore: South Sámi

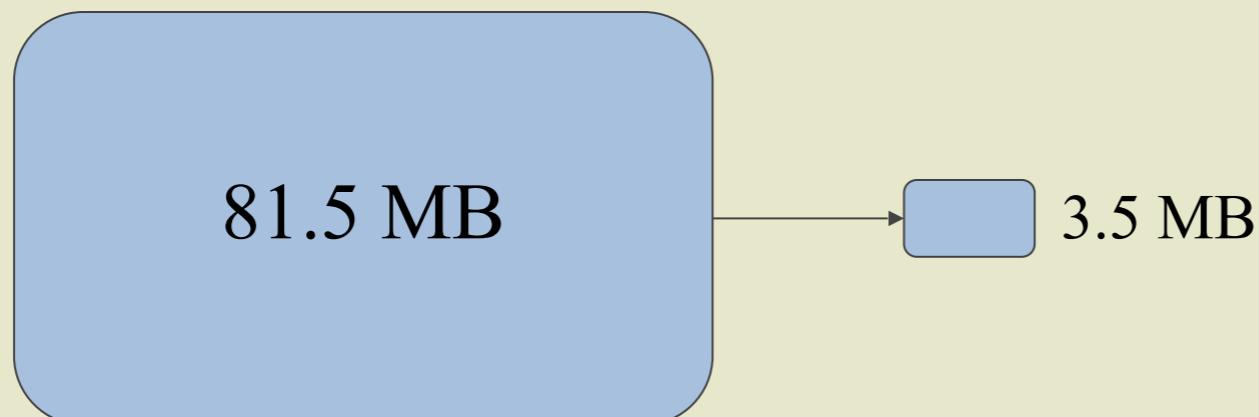
Lemma extraction

```
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>dihete</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>jääktan</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">verbgren. · · · </c></i></small><small><kref>vuejehten</kref> · </small>
1255064 </indentIt></ln><span><small><small>
1255065 </small></small><indentIt>Eksempler: · · · </indentIt>
1255066 · · · <small>vuejiehtidh· birketjem</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>vuejemelohkeht&ja</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>·<i>ulikest.</i></small></small><span>·<small><c·c="dimgray"></c></small><span><bf>å·</bf>jage</span>·<small><small></small></small>·<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><small><small></small><small><small>Eksempler:</small><small>vuejiehtidh·bírhketjem</small><i><small><c·c="dimgray">drive·den·lille·flokk</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>vuejieht&ja</k></head><body><small><i>subst.</i></small><span>·<small><c·c="dimgray"></c></small><span>en·som·driver·flokk</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 · birketjem</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-å·drive·en·flokk | | | | |
| vuejiehtidh·bírhketjem——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 birketjem——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.

verb

reclassified

| | | | | | |
|--------|----------------------|--------|----------------------|--------|----------------------|
| 242514 | mn | 242527 | mn | 242552 | mn |
| 242515 | dn | 242528 | dn | 242553 | dn |
| 242516 | sma-000 | 242529 | sma-000 | 242554 | sma-000 |
| 242517 | vuejemelohkehtæija | 242530 | vuejiehtidh | 242555 | vuejiehtidh |
| 242518 | dcs2 | 242531 | dcs2 | 242556 | dcs2 |
| 242519 | art-303 | 242532 | art-303 | 242557 | art-303 |
| 242520 | PartOfSpeechProperty | 242533 | PartOfSpeechProperty | 242558 | PartOfSpeechProperty |
| 242521 | art-303 | 242534 | art-303 | 242559 | art-303 |
| 242522 | Noun | 242535 | Verbal | 242560 | Verbal |
| 242523 | dn | 242536 | dpp | 242561 | dpp |
| 242524 | nob-000 | 242537 | art-303 | 242562 | art-303 |
| 242525 | kjørelærer | 242538 | Class | 242563 | Class |
| | | 242539 | ulikest | 242564 | ulikest |
| | | 242540 | df | 242565 | df |
| | | 242541 | nob-000 | 242566 | nob-000 |
| | | 242542 | (å)·jage | 242567 | (å)·drive·en·flokk |
| | | 242543 | dn | 242568 | df |
| | | 242544 | nob-000 | 242569 | nob-000 |
| | | 242545 | jage | 242570 | drive·en·flokk |
| | | 242546 | dcs2 | | |
| | | 242547 | art-303 | | |
| | | 242548 | PartOfSpeechProperty | | |
| | | 242549 | art-303 | | |
| | | 242550 | Verbal | | |

Results

- Demonstrated use cases
 - PanImages

The screenshot shows the PanImages website interface. At the top, there is a logo of the Earth and the text "PanImages Cross-Lingual Image Search". Below the logo, a search bar contains the word "spring". To the right of the search bar are "Translate" and "Show Images" buttons, and a dropdown menu set to "Searching in: English". Below the search bar, there is a link "Useful Instructions and Tips" and an email address "Please send feedback to panimages@cs.washington.edu". A "Login" and "Register" button are located at the top right.

On the left, a sidebar titled "Select a meaning:" lists various meanings of "spring":

1. spring (noun): season (122 Translations)
2. spring (verb): jump or leap (93 Translations)
3. spring (noun): water source (81 Translations)
4. spring (noun): device made of flexible material (38 Translations)
5. spring (verb): start to exist (28 Translations)
6. spring (noun): higher-than-average tide (3 Translations)
7. spring (noun): rope on a boat (2 Translations)
8. spring (noun): the source of an action (2 Translations)
9. fountain (noun) (121 Translations)
10. leap (verb): to jump from one location to another (147 Translations)

Other translations (See more meanings) and (Edit meanings) links are also present.

The main content area displays "Meaning 1 of 42: spring (noun): season Languages with many images (See more...)" and "Edit translations". It lists translations for various languages:

| Language | Translation |
|--------------------|---------------|
| Breton | nevezamzerioù |
| | nevezamzer |
| | nevez-amzer |
| Bulgarian | първ |
| Catalan | primavera |
| Chinese (Mandarin) | 春天 |
| Croatian | proljeće |
| Czech | jaro |
| Danish | forår |
| | forårs |
| Dutch | lente |
| | voorjaar |
| English | spring |

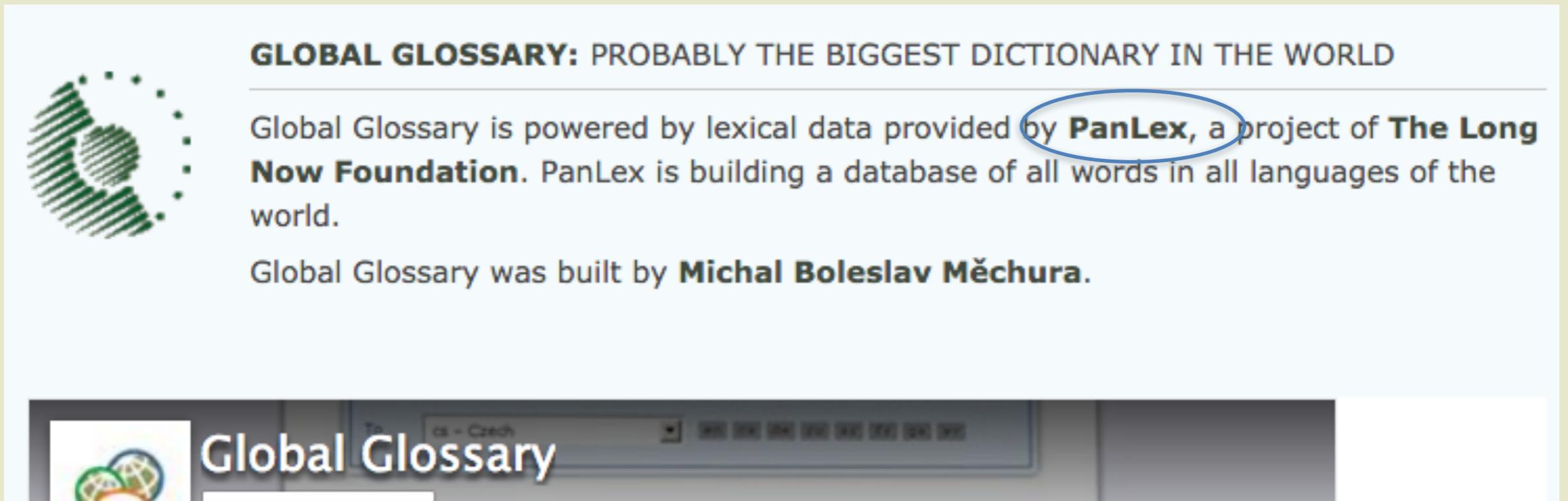
The screenshot shows search results for "spring" on Flickr and Google.

Flickr: The page displays the Flickr logo and a message stating "We're sorry, Flickr doesn't allow embedding within frames. If you'd like to view this content, [please click here](#)." There is a "Return to PanImages" link and options to "View results only from Flickr" or "View results only from Google".

Google: The search results page for "spring" shows several image thumbnails. One thumbnail for a yellow and white flower is labeled "Gul og hvid anemone. 400 x 400 - 19kb - jpg skoven-i-skolen.dk Find lignende billeder". Another thumbnail for a white flower is labeled "Forår blomster. 300 x 448 - 47kb - jpg digitalphoto.pl Find lignende billeder". Other results include "Forside 107231" and "Forside".

Results

- Demonstrated use cases
 - Global Glossary



GLOBAL GLOSSARY: PROBABLY THE BIGGEST DICTIONARY IN THE WORLD

Global Glossary is powered by lexical data provided by **PanLex**, a project of **The Long Now Foundation**. PanLex is building a database of all words in all languages of the world.

Global Glossary was built by **Michal Boleslav Měchura**.

The screenshot shows the Global Glossary homepage. On the left is the PanLex logo, which consists of a green globe icon with white lines and dots. To the right of the logo, the text "GLOBAL GLOSSARY" is displayed in large, bold, blue letters, followed by the subtitle "PROBABLY THE BIGGEST DICTIONARY IN THE WORLD" in smaller, regular blue letters. Below this, there is a paragraph of text about the project's power source. At the bottom of the page, there is a navigation bar with a "To" dropdown menu set to "cs - Czech" and several other language options like "en - English", "de - German", etc. There is also a search bar and some other site navigation links.

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

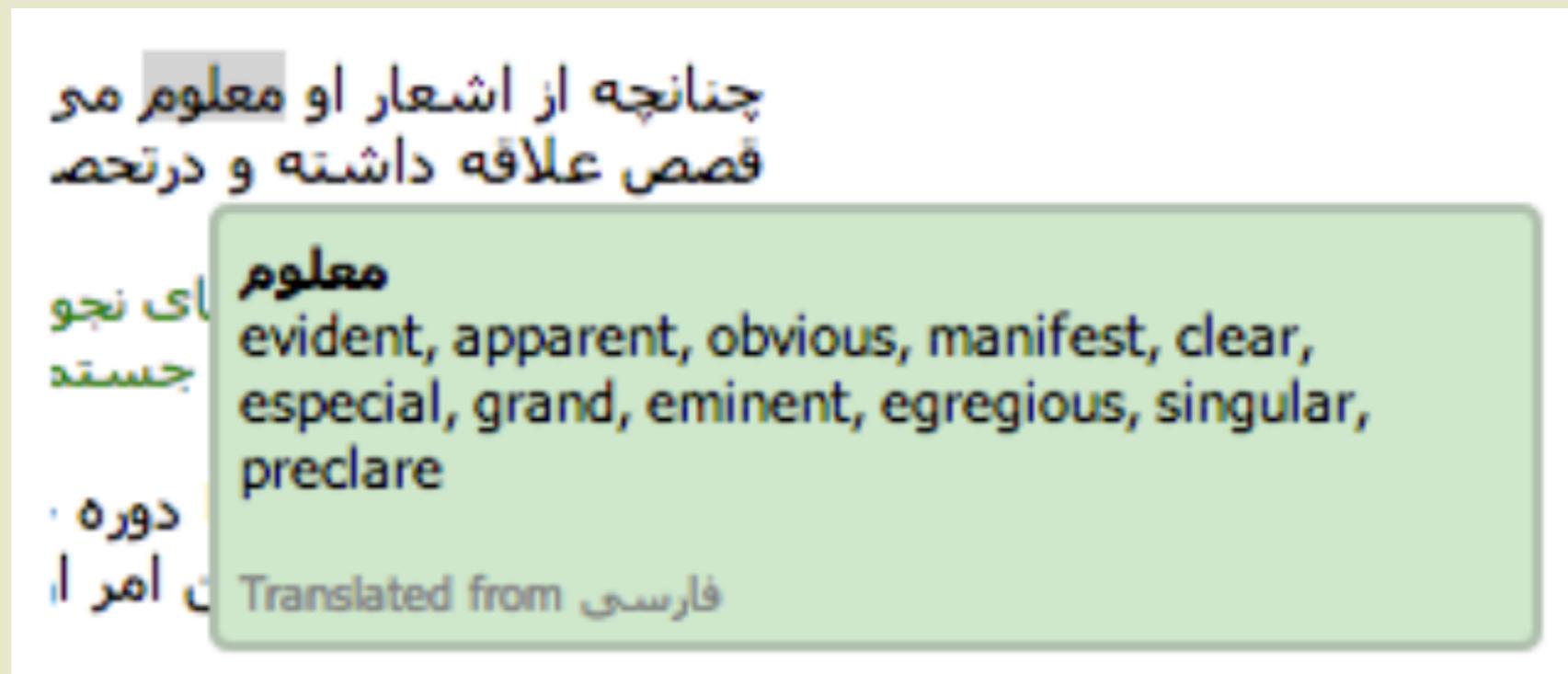
- TeraDict

[Live demo](#)

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

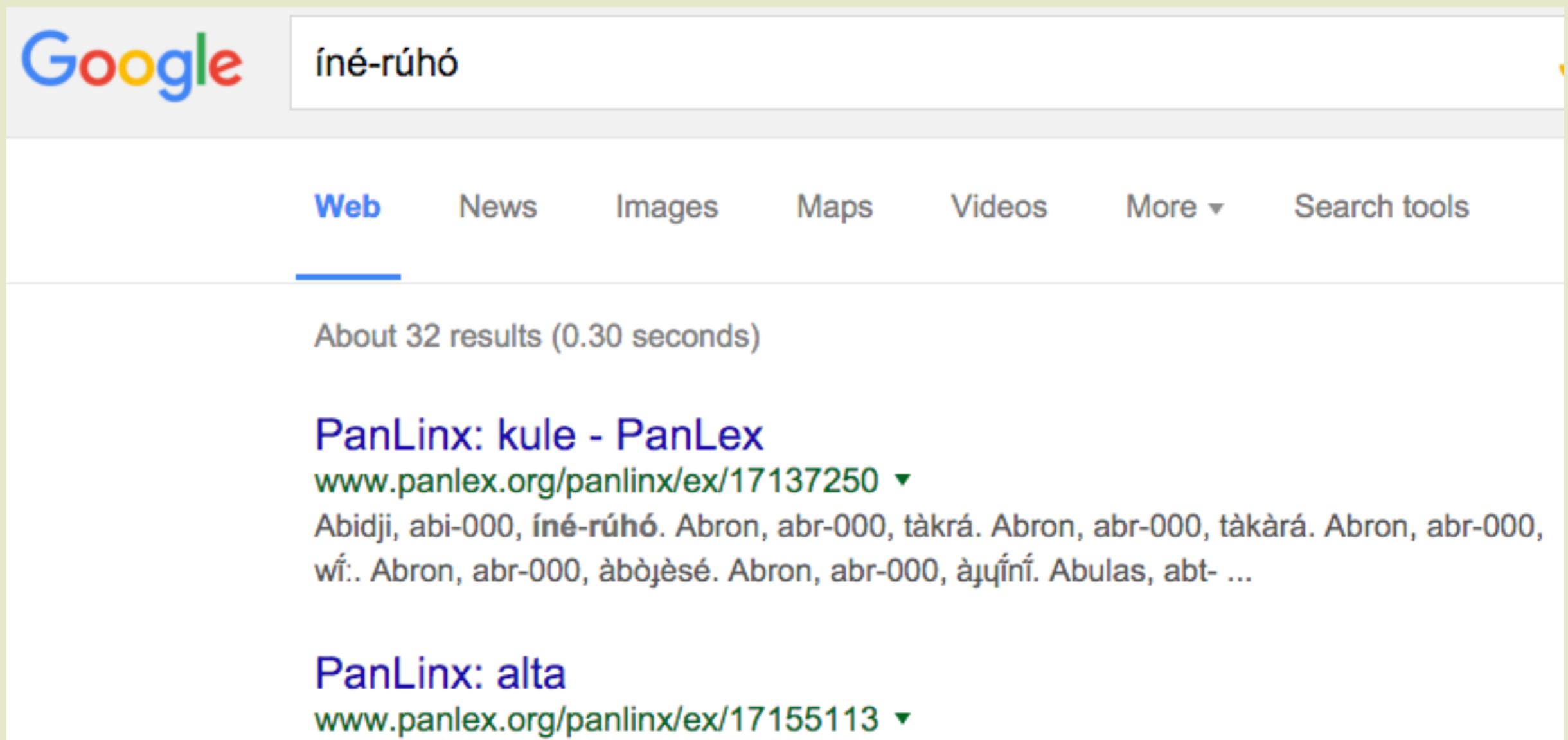
Access

- PanLex Translator (Chrome extension)



Access

- PanLinx



A screenshot of a Google search results page. The search query "íné-rúhó" is entered in the search bar. The results are filtered by "Web". There are approximately 32 results found in 0.30 seconds. The first result is a link to "PanLinx: kule - PanLex" with the URL www.panlex.org/panlinux/ex/17137250. The snippet for this result includes the text "íné-rúhó. Abron, abr-000, tàkrá. Abron, abr-000, tàkàrá. Abron, abr-000, wí:. Abron, abr-000, àbòjèsé. Abron, abr-000, àjúñí. Abulas, abt- ...". The second result is a link to "PanLinx: alta" with the URL www.panlex.org/panlinux/ex/17155113.

Access

- PanLem

PanLex 4.0

Yoruba

seséyédékun

ìtumò

ení — rífojú ríriíran
gbogbo — rífojú ríriíran

ìsọlísóró

rífojú ríriíran

ìtumó oríkì

rífojú ríriíran

èdè — yàn şà

èdè

rífojú ríriíran
laáisce àiní

eníkan

rífojú ríriíran

மொழி — தேர்ந்தெடு

மொழி

காண்
அவசியமாக இரு

ஆள்

காண்
புதிய

மூலம்

பெயர் — காண்
எண் — காண்
பெயர் — ஆண்டு — தலைப்பு -
அமைப்பு

Í ìtumóàtúmóímó

Tamil

PanLex 4.0

orú
iye
orú
rífo
ìtum

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 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: Creative 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 with

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?