Designing a panlingual dictionary

Jonathan Pool • Susan Colowick • Laura Welcher

The Long Now Foundation

PanLex

http://panlex.org

36th Internationalization & Unicode Conference

23 October 02012

Summary

- Introduction
  1. Objective
  2. Team
  3. Strategy
  4. PanLex metrics
  5. Sources
- Design
  1. Schema
  2. Example
  3. Directionality
- Standardization
  1. Language varieties
  2. Character encodings
  3. Normalization forms
  4. Character admissibility
  5. Lexemes
  6. Lemmas
  7. Lexical classification
- Opportunities
- Try it

Introduction

1. Objective

Look up *any* word in *any* language.

Cusco Quechua: pinchikilla

Get its translation(s) into *any* other language.

Nepali: ?

Introduction

2. Team

02005–02009: University of Washington, Turing Center

“TransGraph”
“PanDictionary”
“PanImages”
“Panlingual Translator”
“Panlingual Mail”
“Lemuel”

- Oren Etzioni
- Katherine Everett
- Christopher Lim
- Mausam
- Kobi Reiter
- Marcus Sammer
- Michael Schmitz
- Michael Skinner
- Stephen Soderland
- Timothy Baldwin
- Jonathan Pool
- Susan M. Colowick
- Janara Christensen
- Daniel S. Weld
- Jeff Bilmes
- Katrin Kirchhoff
- Bo Qin

02010–02011: Utilika Foundation

“PanLex”

- Jonathan Pool
- Susan M. Colowick
- Timothy Usher
- Christa Mabee
- Michael Goodman
- David Howcroft
- Miranda Taylor

02012–: The Long Now Foundation

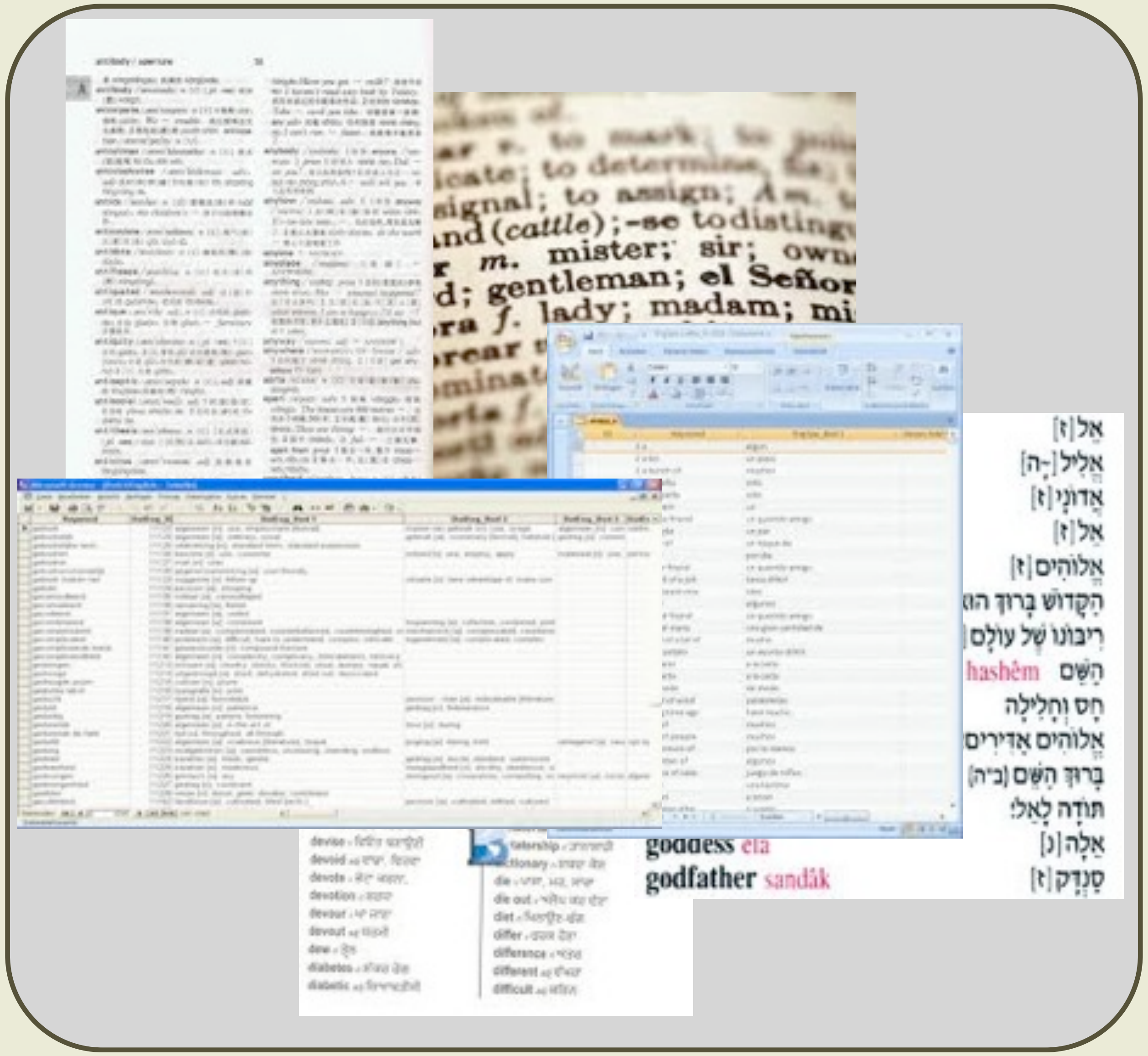
“PanLex”

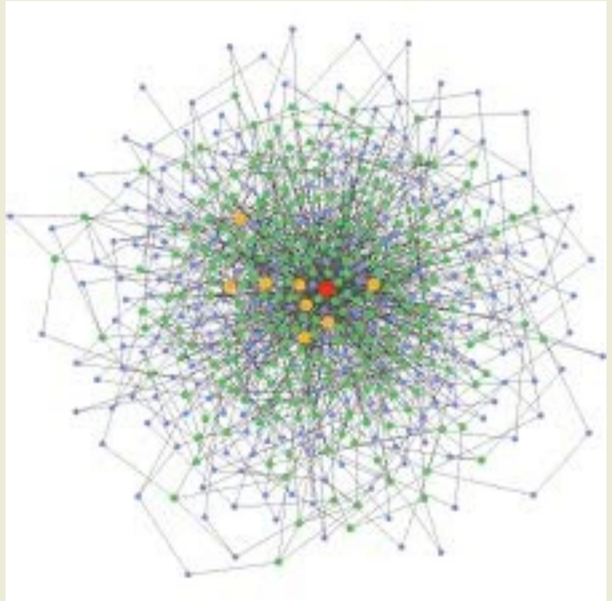
- Jonathan Pool
- Susan M. Colowick
- Andréa Davis
- Laura Welcher
- Ben Keating
- Kurt Bollacker
- Emily Bender
- Steven Bird

Introduction

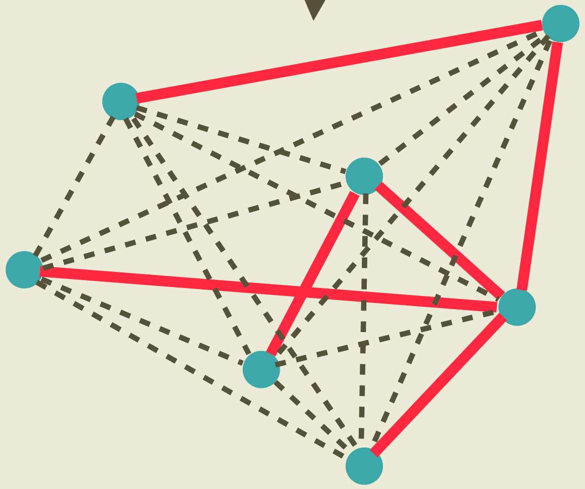
3. Strategy

a. Combine *all* known lexical translations into a database.


to determine
signal; to assign
and (cattle); -se to disting
m. mister; sir; own
d; gentleman; el Señor
ra f. lady; madam; mi
hashêm הַשֵּׁם
goddess ela
godfather sandâk



b. Fill in the gaps with automated inference.

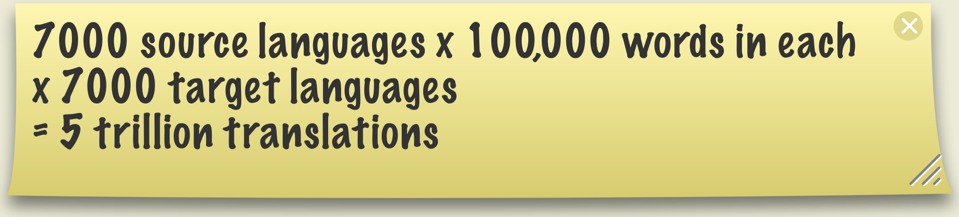


Introduction

4. PanLex metrics

- 18 million expressions (words or phrases).
- 6,900 language varieties.
- 1,400 sources consulted.
- 460 million translations (expression pairs).

Goal: *trillions* of translations



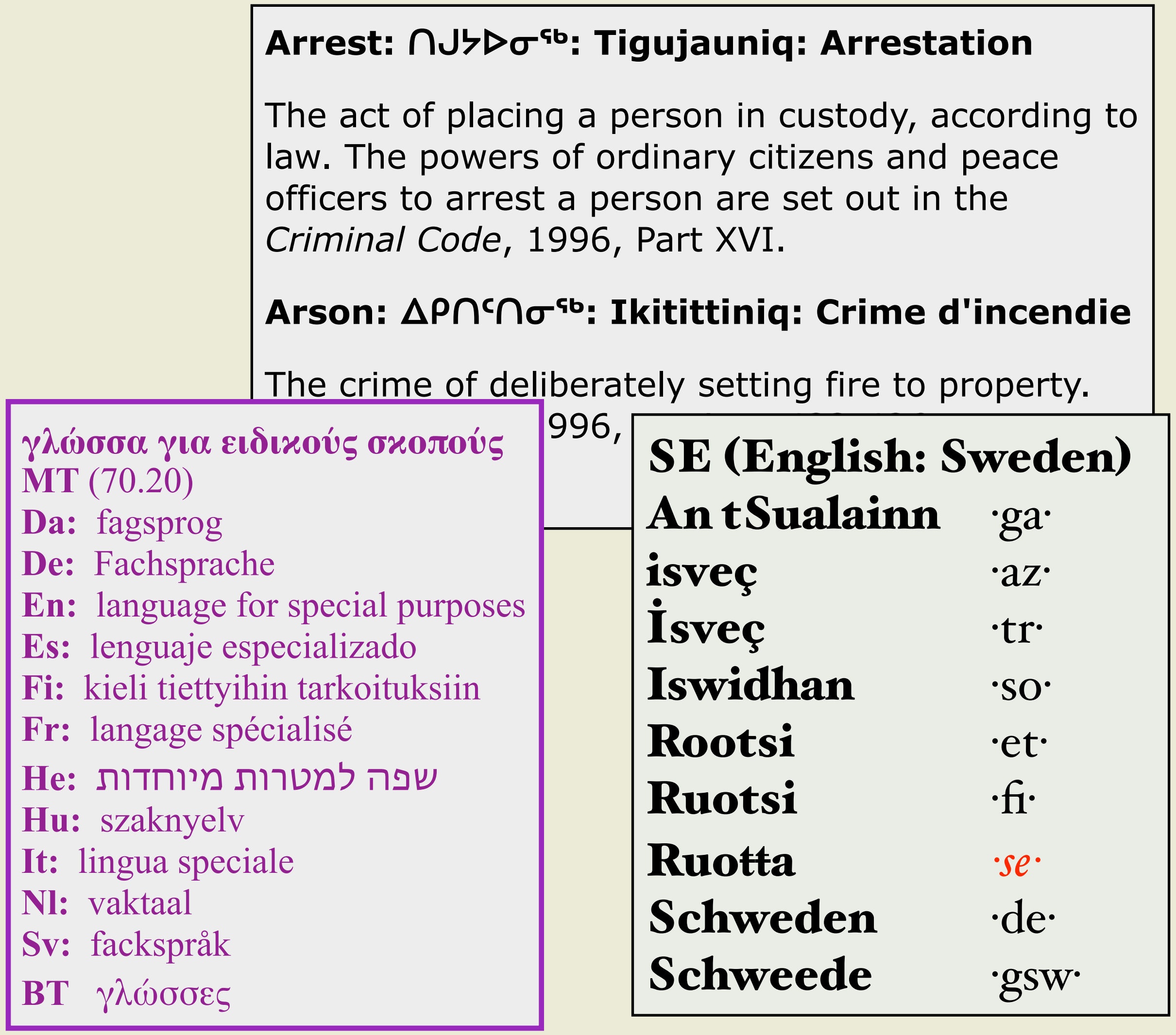
7000 source languages x 100,000 words in each
x 7000 target languages
= 5 trillion translations

Introduction

5. Sources

- Monolingual dictionaries
- Bilingual dictionaries
- Multilingual dictionaries
- Wiktionaries
- Glossaries
- Wordlists
- Terminologies
- Wordnets
- Thesauri
- Standards
- Locale databases
- Vocabulary databases
- Locale databases
- Subject heading lists

E.g., CLDR



Arrest: ᑎᒍᔭᐅᓂᖅ: Tigujauniq: Arrestation

The act of placing a person in custody, according to law. The powers of ordinary citizens and peace officers to arrest a person are set out in the *Criminal Code*, 1996, Part XVI.

Arson: ᐃᑭᑎᑦᑎᓂᖅ: Ikitittiniq: Crime d'incendie

The crime of deliberately setting fire to property.

996,

γλώσσα για ειδικούς σκοπούς

MT (70.20)

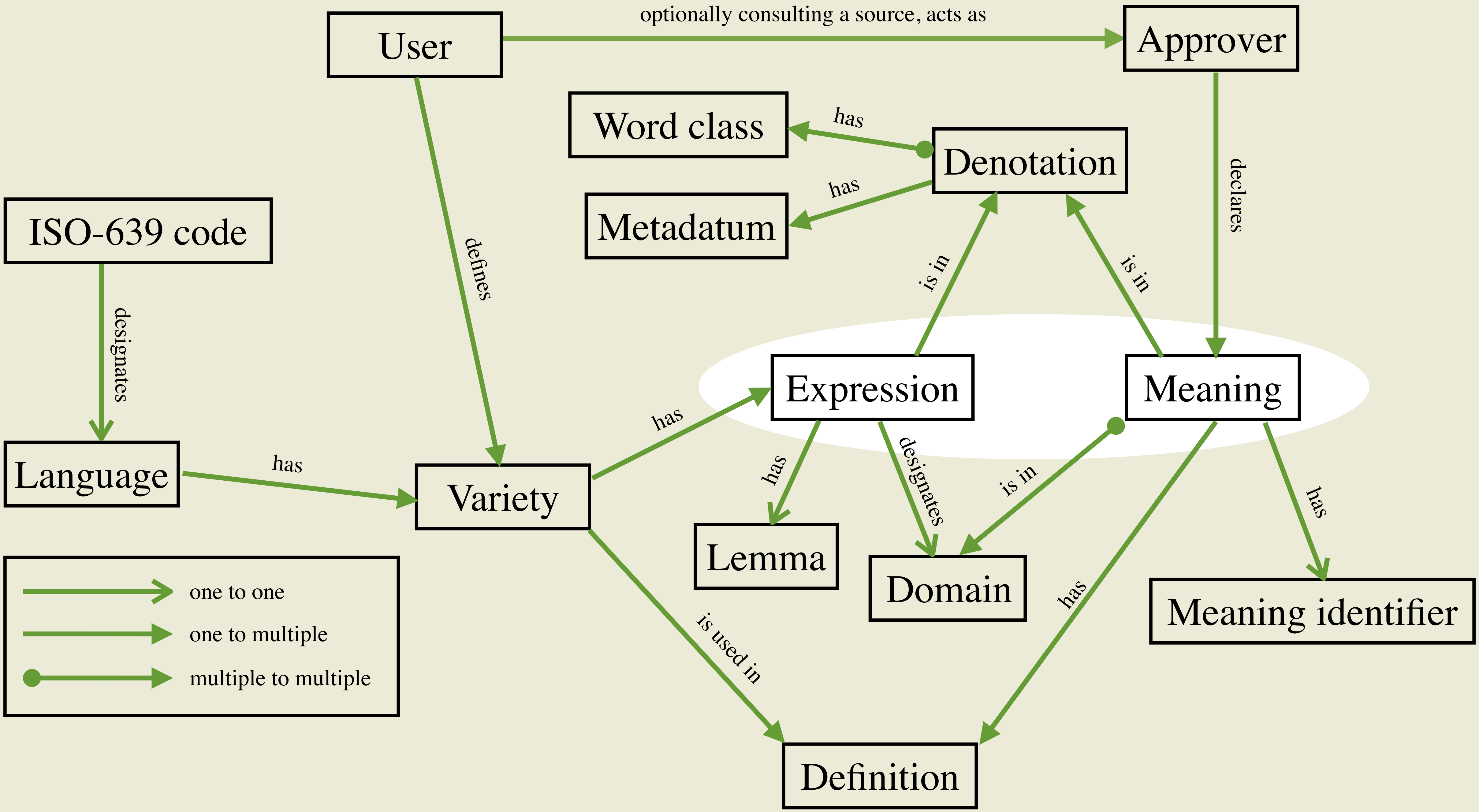
Da:	fagsprog
De:	Fachsprache
En:	language for special purposes
Es:	lenguaje especializado
Fi:	kieli tiettyihin tarkoituksiin
Fr:	langage spécialisé
He:	שפה למטרות מיוחדות
Hu:	szaknyelv
It:	lingua speciale
Nl:	vaktaal
Sv:	fackspråk
BT	γλώσσες

SE (English: Sweden)

An tSualainn	·ga·
isveç	·az·
İsveç	·tr·
Iswidhan	·so·
Rootsi	·et·
Ruotsi	·fi·
Ruotta	*·se·*
Schweden	·de·
Schweede	·gsw·

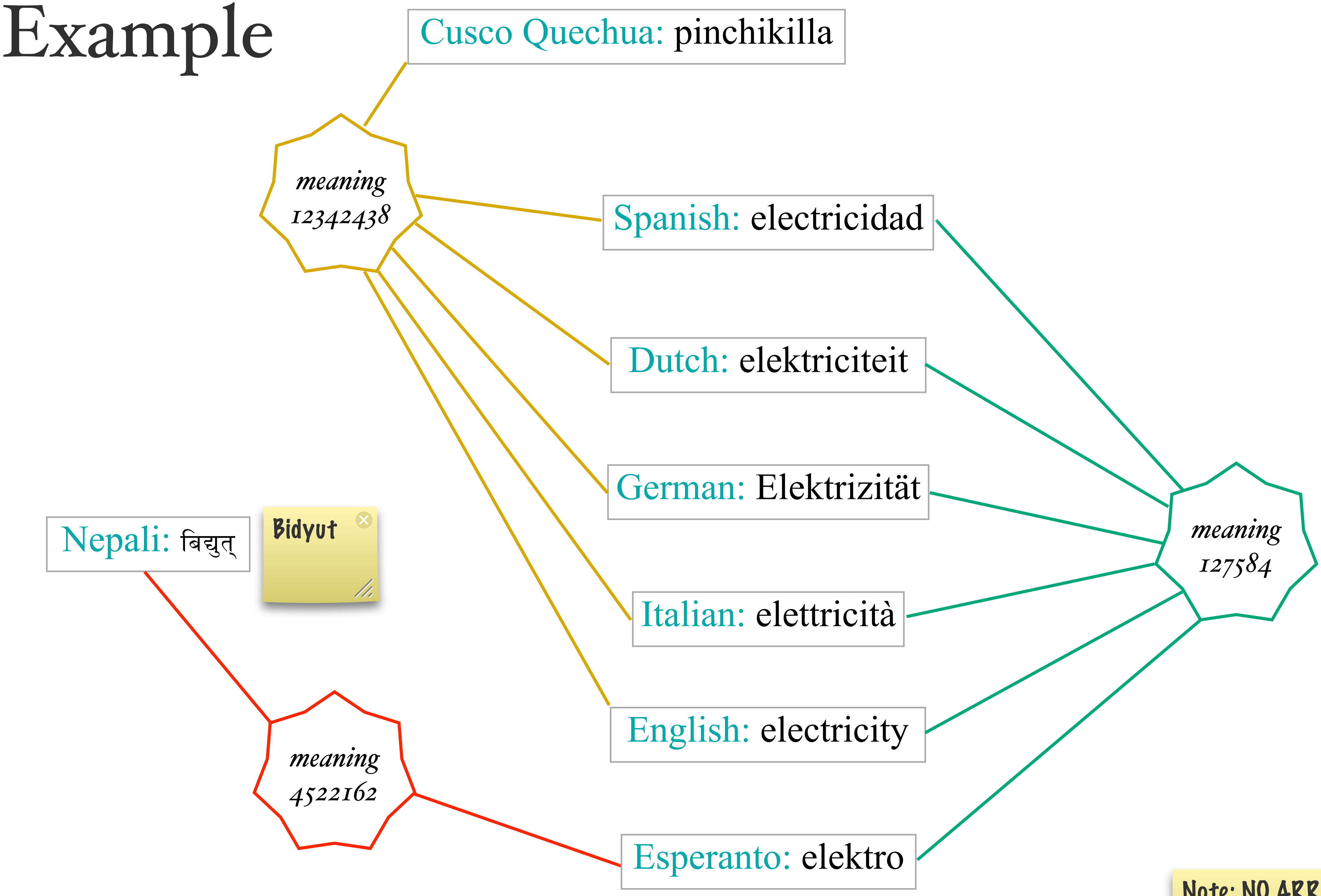
Design

1. Schema



Design

2. Example



Cusco Quechua: pinchikilla

meaning 12342438

Spanish: electricidad

Dutch: elektriciteit

German: Elektrizität

Italian: elettricità

English: electricity

Esperanto: elektro

meaning 127584

Nepali: बिद्युत्

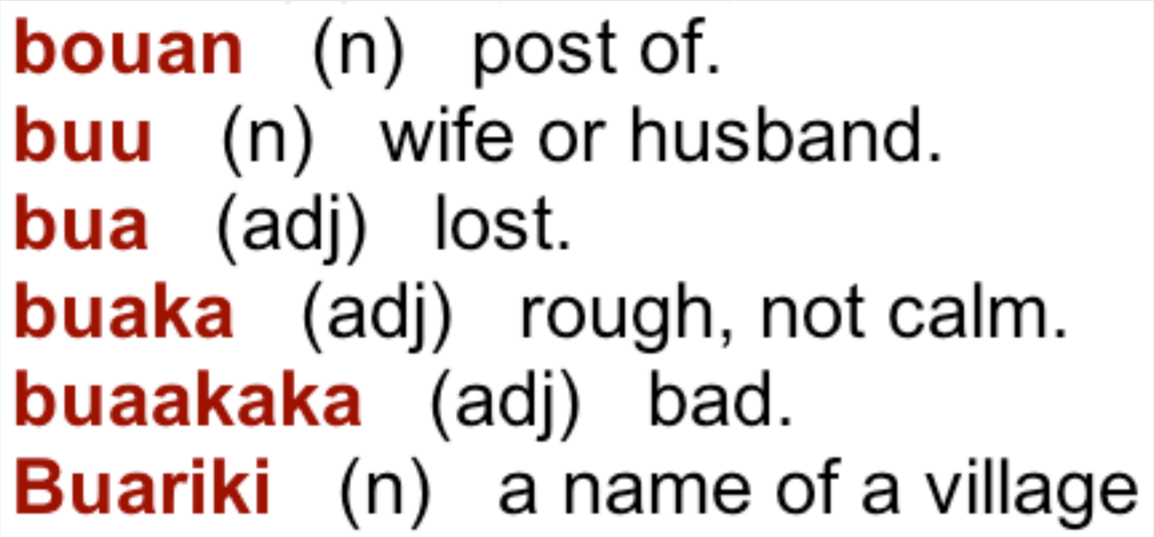
Bidyut

meaning 4522162

Note: NO ARROWS!
Could invert search.

Design

3. Directionality

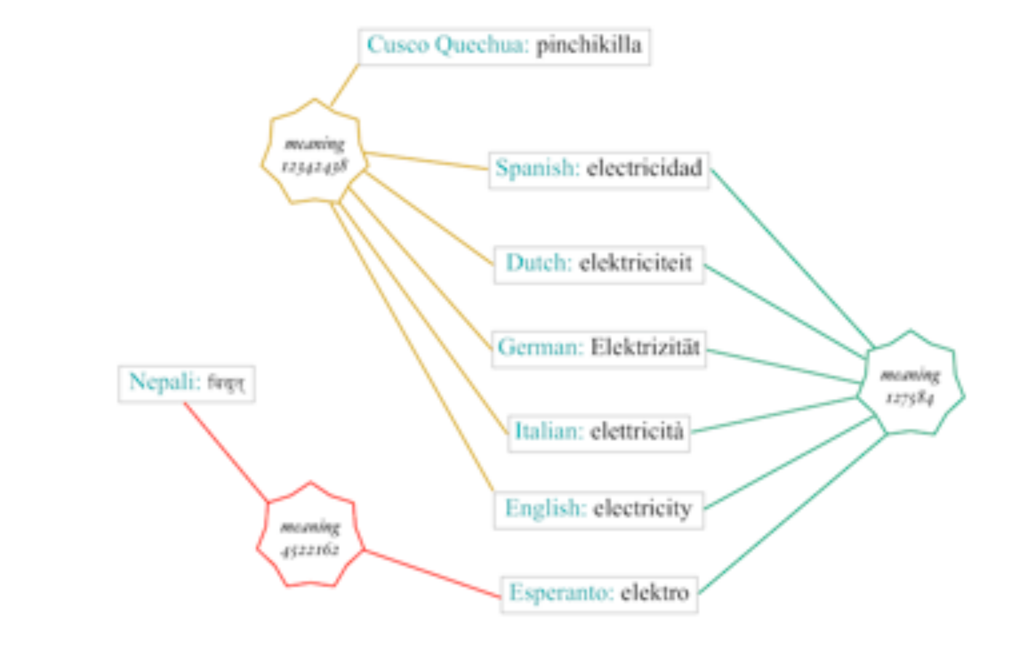


bouan	(n)	post of.
buu	(n)	wife or husband.
bua	(adj)	lost.
buaka	(adj)	rough, not calm.
buaakaka	(adj)	bad.
Buariki	(n)	a name of a village

Dictionary (typically)

Source expressions translated into and/or explained in *target* languages. Directional.

vs.



PanLex

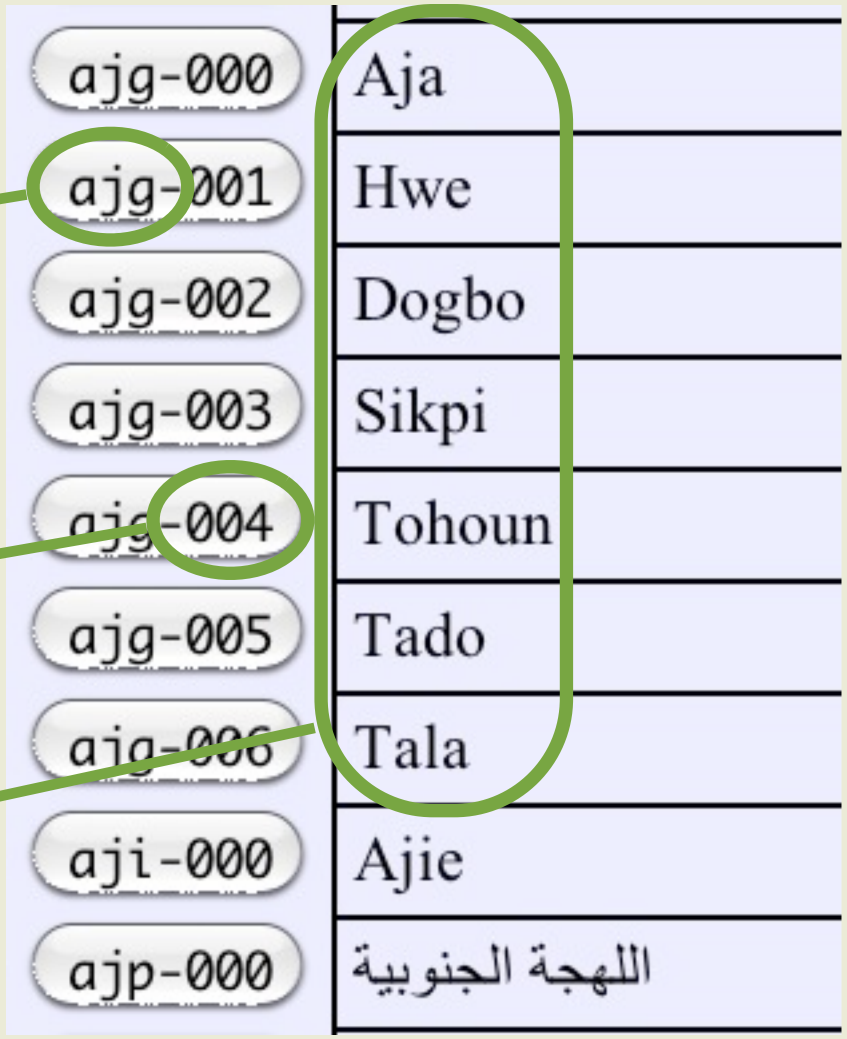
Expressions sharing meanings are translations of *each other*. **Non**directional.

Standardization

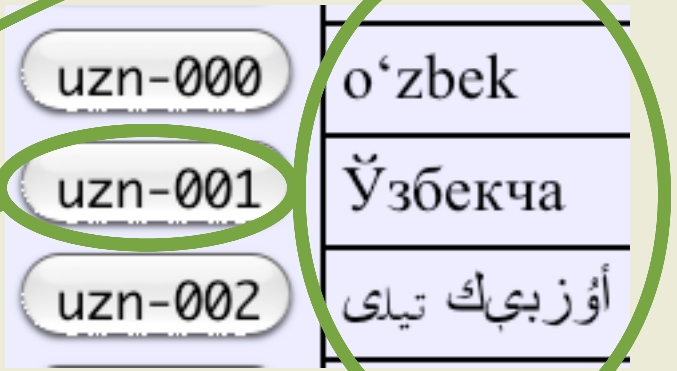
Dialects of Aja, a language of Benin and Togo

1. Language varieties

- "Languages" identified with ISO 639-2, 3, 5 alpha-3 codes.
- "Varieties" identified with integers (for free extensibility).
- Dialectal, standard, controlled, script varieties.



ajg-000	Aja
ajg-001	Hwe
ajg-002	Dogbo
ajg-003	Sikpi
ajg-004	Tohoun
ajg-005	Tado
ajg-006	Tala
aji-000	Ajie
ajp-000	اللهجة الجنوبية



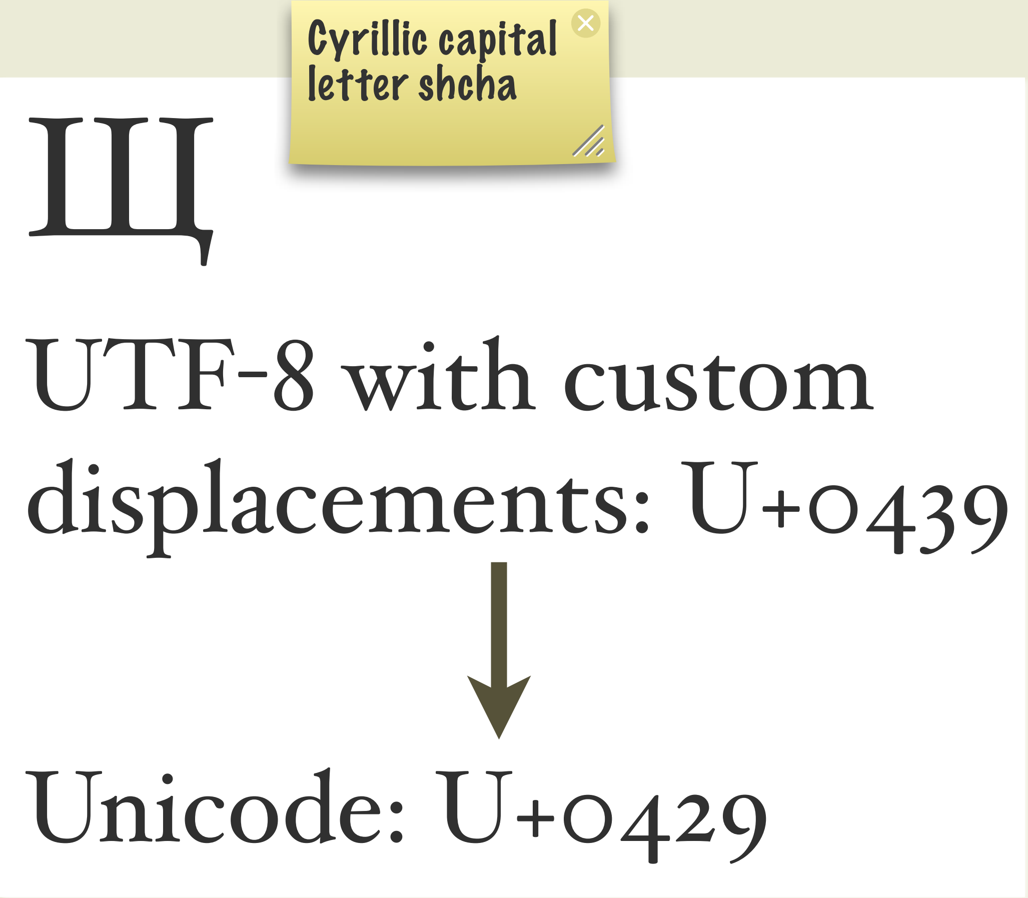
uzn-000	o‘zbek
uzn-001	Ўзбекча
uzn-002	أۇزبېك تیلی

Cf. BCP 47: uz-uzn-Cyrl

Standardization

2. Character encodings

- Unicode.
- UTF-8 encoding form.



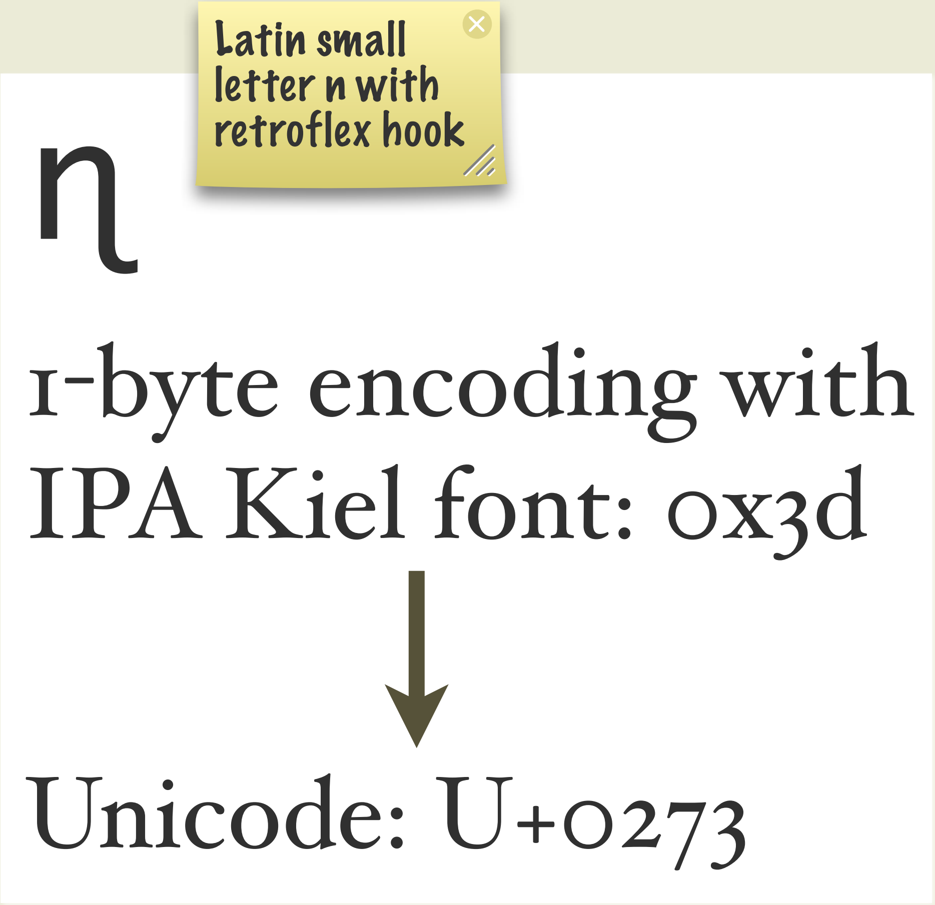
Cyrillic capital letter shcha

Щ

UTF-8 with custom displacements: U+0439

↓

Unicode: U+0429



Latin small letter n with retroflex hook

ɳ

1-byte encoding with IPA Kiel font: 0x3d

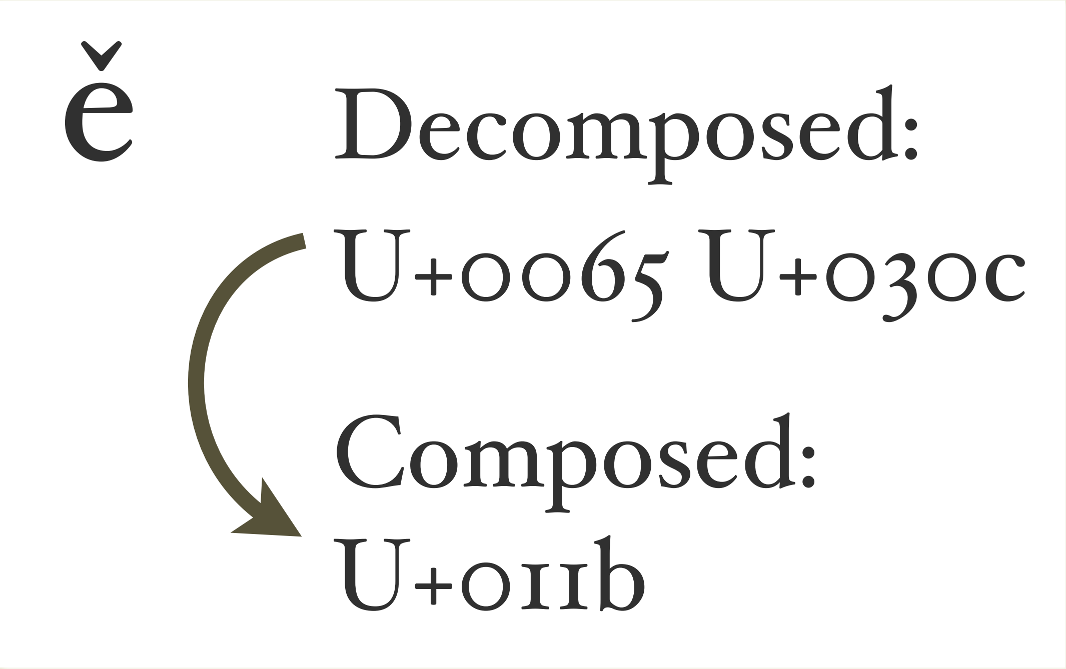
↓

Unicode: U+0273

Standardization

3. Normalization forms

- Normalization Form C (NFC): Canonical decomposition followed by canonical composition.



ě

Decomposed:
U+0065 U+030c

Composed:
U+011b

- NFC leaves visual ambiguities. (Even NFKC would eliminate only 1 of these.)

/	U+002f
／	U+ff0f
⁄	U+2044
╱	U+2571
∕	U+2215

Standardization

4. Character admissibility

- Exclude characters with Other Unicode General Category Properties.

~~SOFT HYPHEN~~

~~U+00ad~~

- Exclude characters with Separator Unicode General Category properties except SPACE.

~~SIX-PER-EM SPACE~~

~~U+2006~~

- Prohibit SPACE at beginnings and ends of strings.

“✗öpmek”

- Prohibit 2 or more consecutive instances of SPACE in any string.

“ztráta ✗barvy”

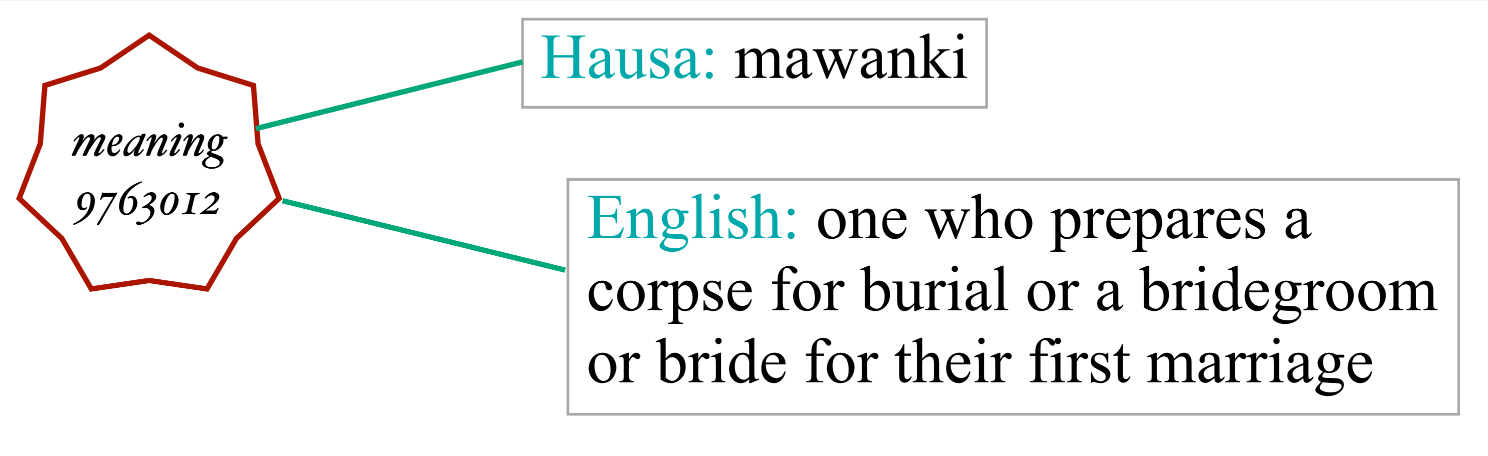
Standardization

5. Lexemes

- “Objective: Look up any **word** in any language ….”
- More precisely: **lexeme**.
- Is a phrase a lexeme?
- Is an inflected form a lexeme?
- If a translation isn’t a lexeme, PanLex editor may:
  - Use it as a definition.
  - Approximate (“undertaker”, “makeup artist”).
  - Coin (“mawanki”).

- “sweet tooth”: yes
- “sweet dessert”: no
- “sweet wine”: ?

- “glasses”: yes
- “statements”: no
- “instructions”: ?


meaning
9763012
Hausa: mawanki
English: one who prepares a corpse for burial or a bridegroom or bride for their first marriage

Standardization

6. Lemmas

- Citation (dictionary lookup) forms of lexemes.
- Standardized, to facilitate connectivity.

U+00f1
Latin small letter
n with tilde

(Abidjan)
U+0027
apostrophe

U+015f
Latin small letter
s with cedilla

English:	Swahili:	Turkmen:	Hebrew:	Romanian:	Esperanto:
~~to~~ share vitamin~~s~~	~~elimisha~~ -elimisha	~~garañky~~ garaňky	~~אביג'אן~~ אביג׳אן	~~cartepoştală~~ carte poștală	~~Kantocigno~~ kantocigno

U+0148
Latin small letter
n with caron

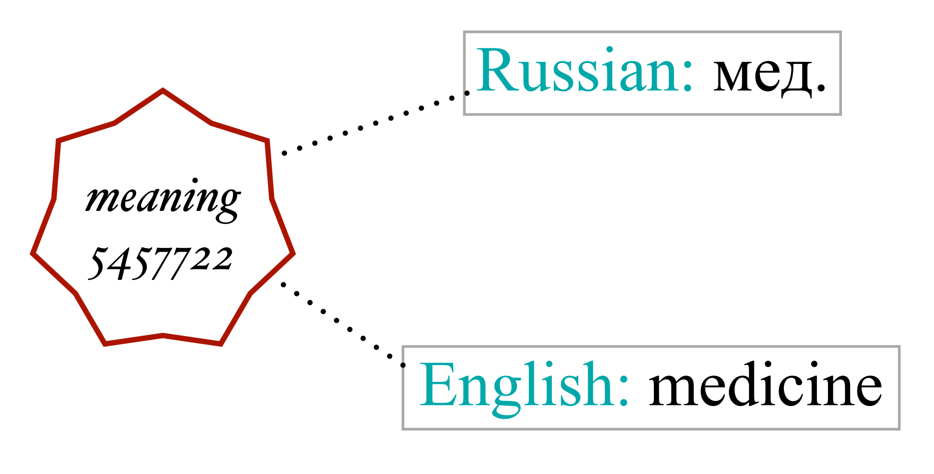
U+05f3
Hebrew
punctuation
geresh

U+0219
Latin small letter
s with comma
below

Standardization

7. Lexical classification

Open set of meaning domains.



Extension of OLIF

Cf. 19 subclasses of GOLD "Part Of Speech Property": Predicator, Functor, Determiner, Noun, ProForm, Classifier, Particle, Quantifier, Expletive, Interjection, InterrogativeOperator, Modal, NegationOperator, Nominal, Participle, Prenoun, Preverb, Substantive, SyntacticArgument

Closed set of 15 word classes.

adjv	adjective
advb	adverb
affx	affix
auxv	auxiliary verb
conj	conjunction
detr	determiner
ijec	interjection
misc	miscellaneous
name	proper noun
noun	noun
post	postposition
prep	preposition
pron	pronoun
verb	verb
vpar	verb particle

Opportunities

- Discover lexical resources.
- Add content.
- Improve quality.
- Refer language experts.
- Create UIs.
- Create APIs.
- Create applications.
- Advise on strategy and tactics.

http://panlex.org/help/

Try it

http://panlex.org/try/

- Easy UI: TeraDict
- Expert self-localizing UI: PanLem
- Search-oriented UI: PanLinx (“waakaa’iganan” @ Google)

More comments/questions?

Info:

http://panlex.org