PanLex

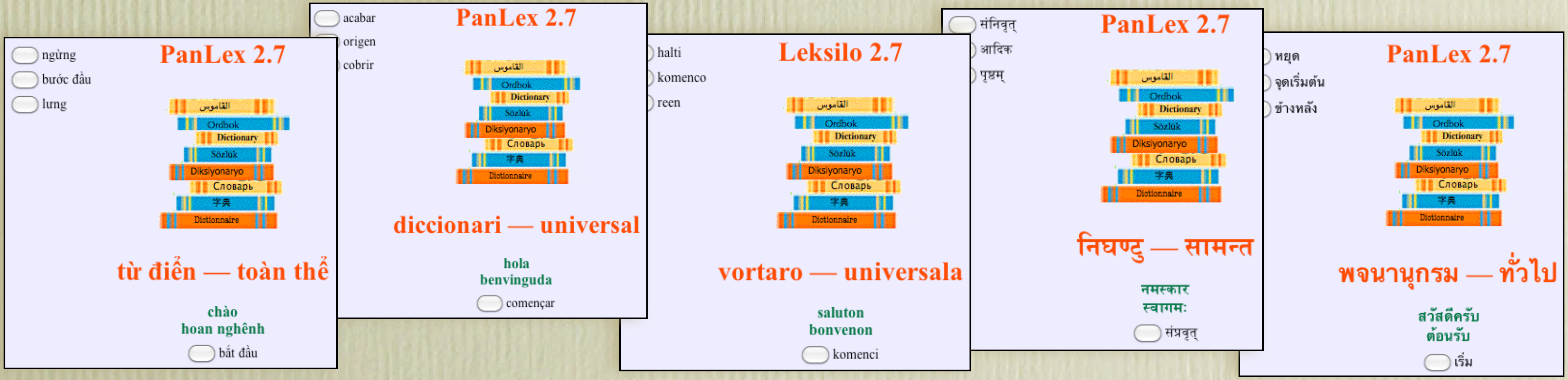
A Panlingual Lexicon

Jonathan Pool & Susan Colowick

Utilika Foundation

DELPH-IN Summit

26 June 2011



Outline

- Purpose
- Related work
- Construction
- Size
- Applications
- Current work
- PanLex and grammar engineering
- Team

Purpose

PanLex aims to become a **panlingual lexical translation resource**.

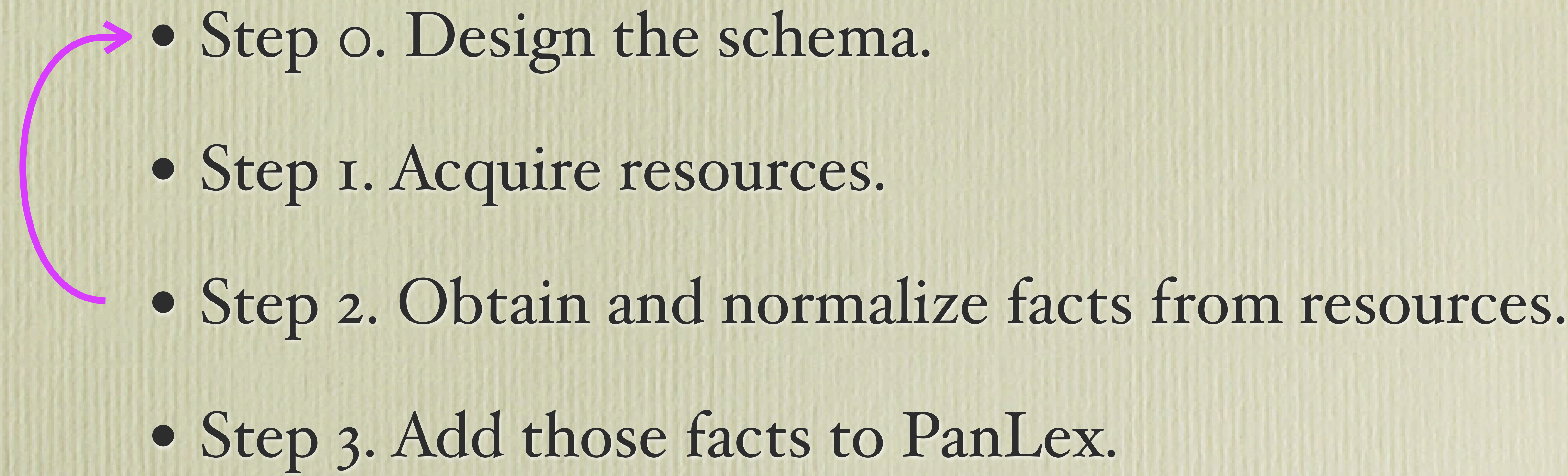
eng-000	English	five
epo-000	Esperanto	kvin
erg-000	Sye	sukrim
erh-000	Eruwa	íì-sòrĩ`
erk-000	Fate	lim
ero-000	Horpa	gwai
ers-000	Ersu	ŋuàr
ers-001	Thochu	wa-re
ers-002	Lyusu	ŋâ
ers-003	Menia	nga
ers-004	Muli	ngo
erw-000	Erokwanas	rim
ese-000	Ese Ejja	me-oe-xi
ese-001	Huarayo	iamatamata
esi-000	Iñupiat	tallimat
esk-000	Iñupiatun	tallimat
esq-000	Huelel	pemaxala
ess-000	Chaplino	tałimat
est-000	Estonian	viis
esu-000	Central Yupik	tałiman
etr-000	Edolo	bi
ett-000	mechl Rasnal	mach
etu-000	Ejagham	é-rôn
etx-000	Aten	wiǀ
etx-001	Niten	wéé
eus-000	euskara	bost
eus-000	euskara	bost
eus-001	Aitzineuskara	*bortz
eus-001	Aitzineuskara	*bortze
eve-000	эвэды торэн	tunŋᵉn
evh-000	Uvbie	i-siorĩ
evn-000	орочон турэн	tunŋa
evn-001	Solon	tongnga
ewe-000	Ɛʋɛgbɛ	ató~
ewo-000	Ewondo	tán
eya-000	Eyak	tcó~i
eyo-000	Keiyo	mú:t
faf-000	Fagani	rima
fag-000	Finungwa	yale yale kobok
fan-000	Pahouin	tan
fao-000	føroyskt	fimm

fay-000	Fars	pänj
fer-000	Feroge	wi
ffm-000	Maasina	joyi
fie-000	Fyer	háwá
fij-000	vosa Vakaviti	e lima na
fin-000	suomi	viisi
fin-000	suomi	viisi
fip-000	Fipa	visaano
fli-000	Fali	kè~r̠ è~w
fmp-000	Fe’fe’	tiǀì
fng-000	Fanagalo	fayif
fni-000	Fanya	luñe
fra-000	français	cinq
fry-000	Frysk	fiif
gil-000	taetae ni Kiribati	nimaua
gla-000	Gàidhlig na h-Alba	còig
gle-000	Gaeilge	cúig
glv-000	chengey Vannin	queig
gug-000	avañe’ẽ	po
heb-000	עברית	חָמֵשׁ
heb-000	עברית	חֲמִשָּׁה
hin-000	हिंदी	पाँच
hrv-000	hrvatski	pet
hun-000	magyar	öt
hye-000	արևելահայերեն	հինգ
ido-000	Ido	kin
ina-000	interlingua	cinque
ind-000	bahasa Indonesia	lima
isl-000	íslenska	fimm
ita-000	italiano	cinque
jbo-000	la lojban.	mu
jpn-000	日本語	五
jpn-000	日本語	五つ
kal-000	kalaallisut	tallimat
kat-000	ქართული	ხუთი
kmr-000	Kurmancî	pênc
kor-000	한국어	다섯
kor-000	한국어	오
kpv-000	коми кыв	вит

Related work

- Dicts.info
- FreeDict
- Freelang
- Global WordNet Association
- Langtolang
- Logos Foundation Dictionary
- OmegaWiki
- OneLook
- Open Dictionary
- The Rosetta Project
- Slovnik
- La Vortaro
- Webster's Online Dictionary
- Wiktionary
- Тезаурус

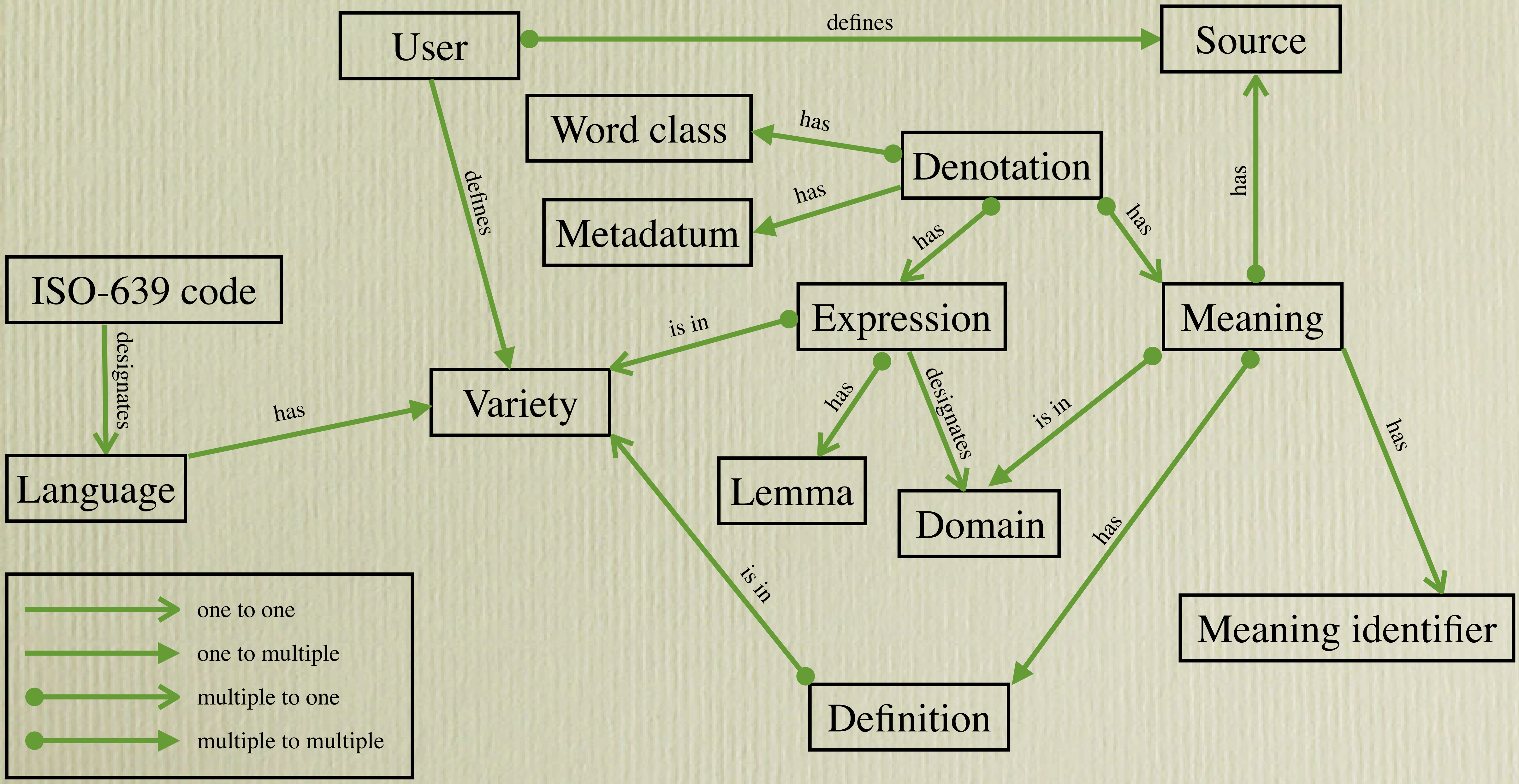
Construction



- Step 0. Design the schema.
- Step 1. Acquire resources.
- Step 2. Obtain and normalize facts from resources.
- Step 3. Add those facts to PanLex.

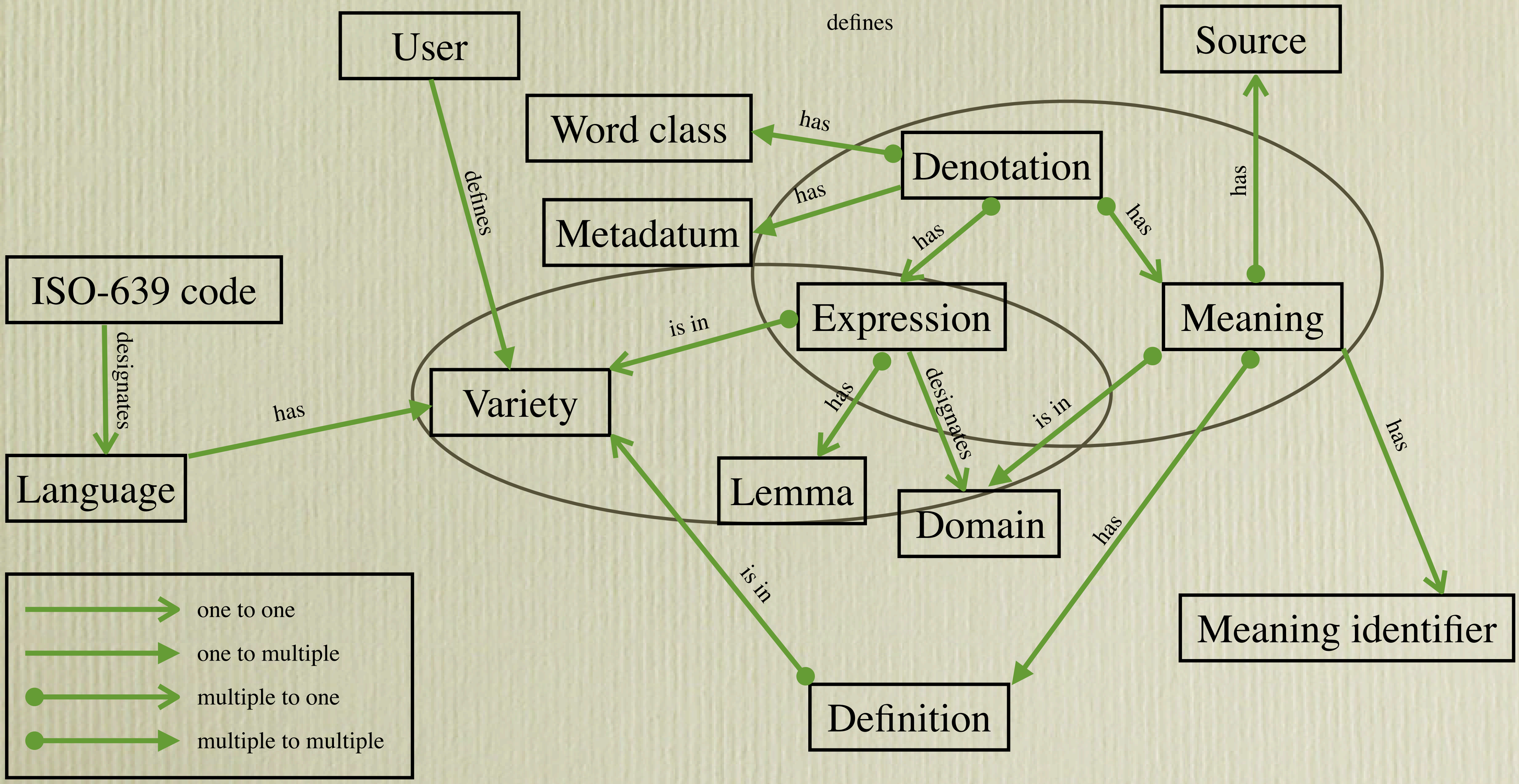
Construction

Step 0. Design the schema.



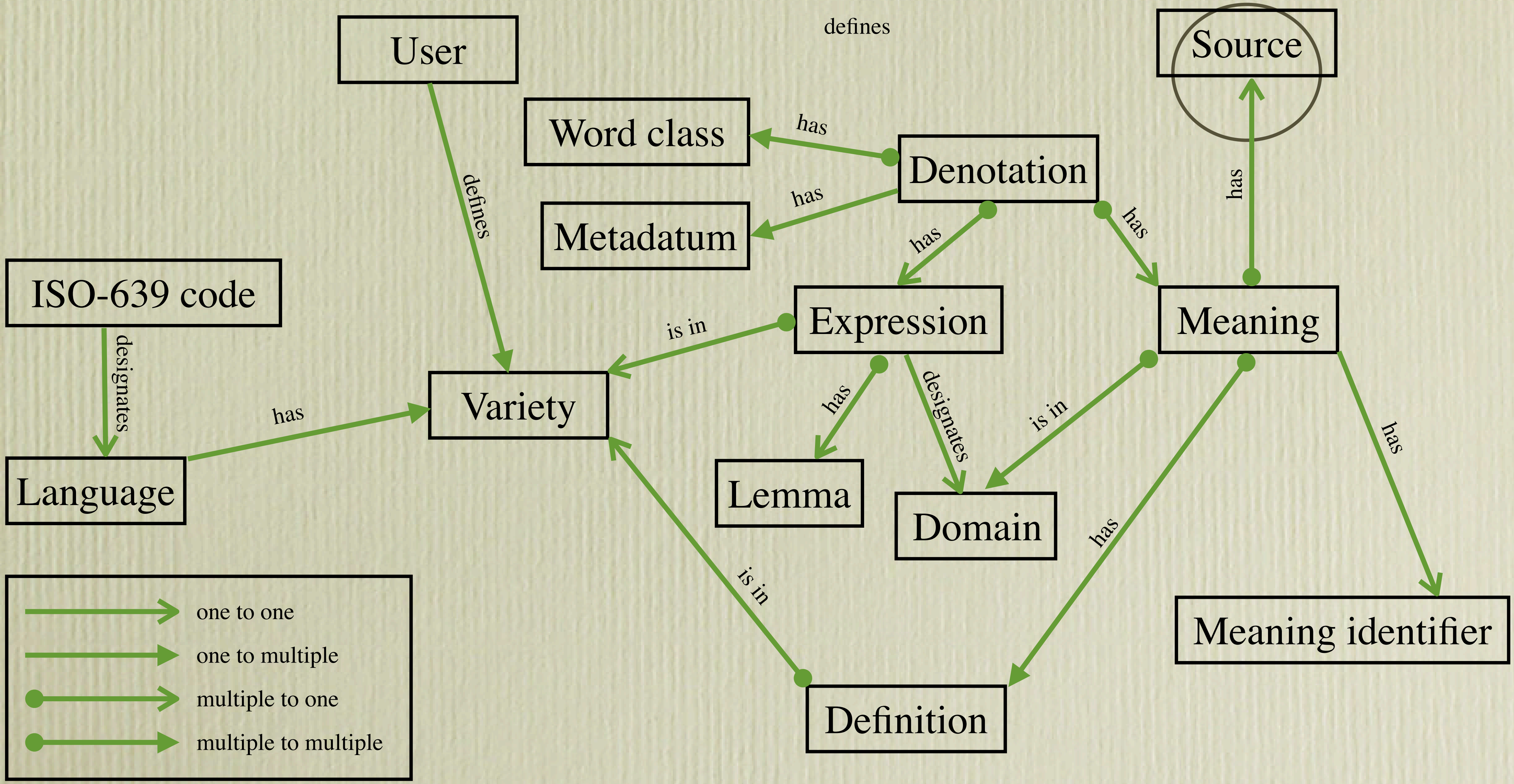
Construction

Step 0. Design the schema.



Construction

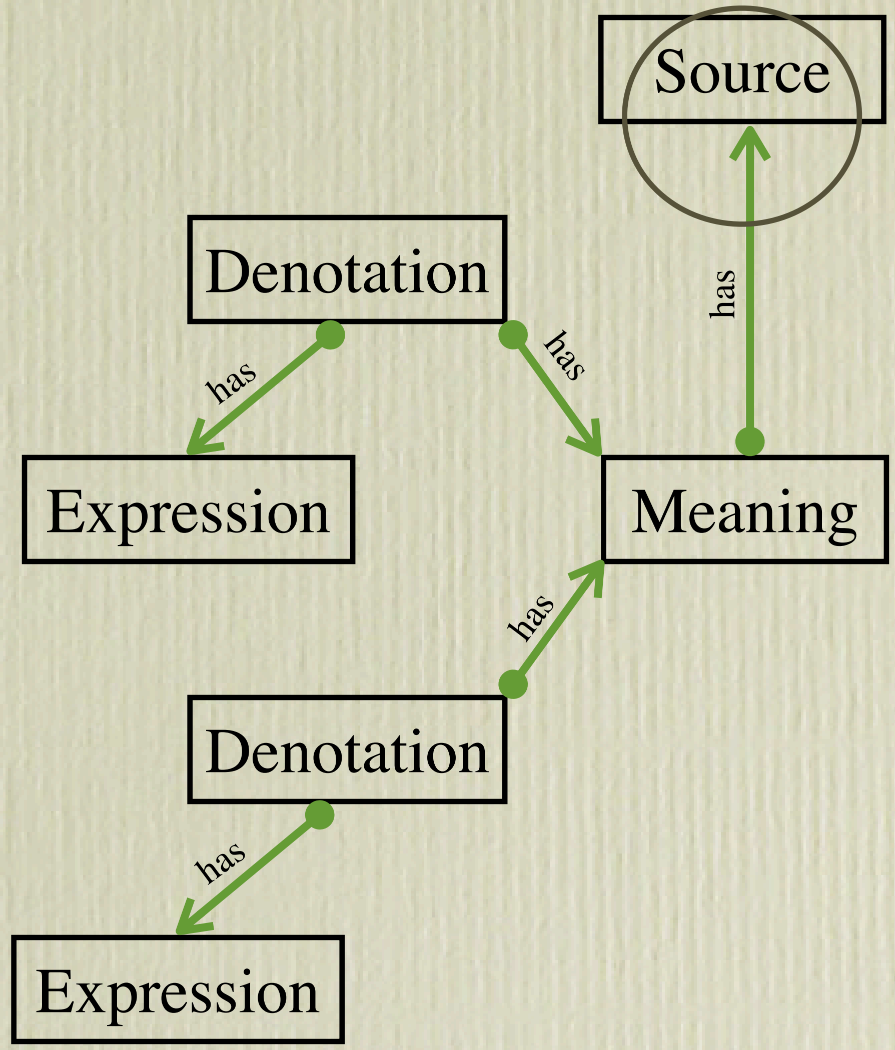
Step 0. Design the schema.

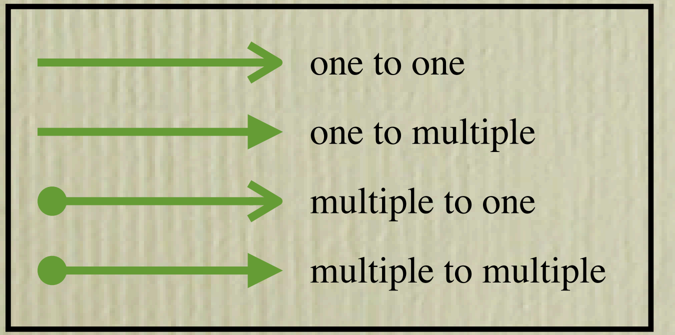


Construction

Step 0. Design the schema.

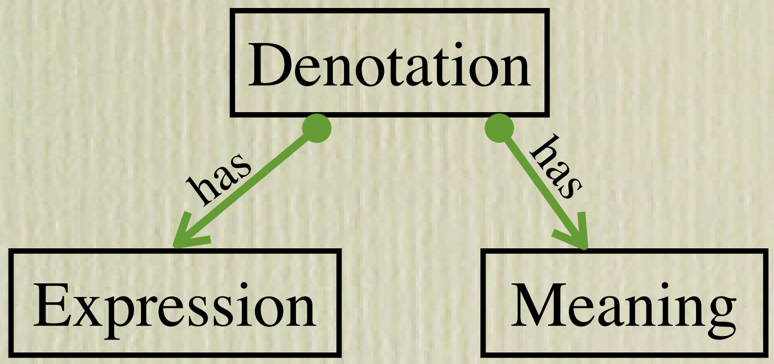
Translation





Construction

Step 0. Design the schema.



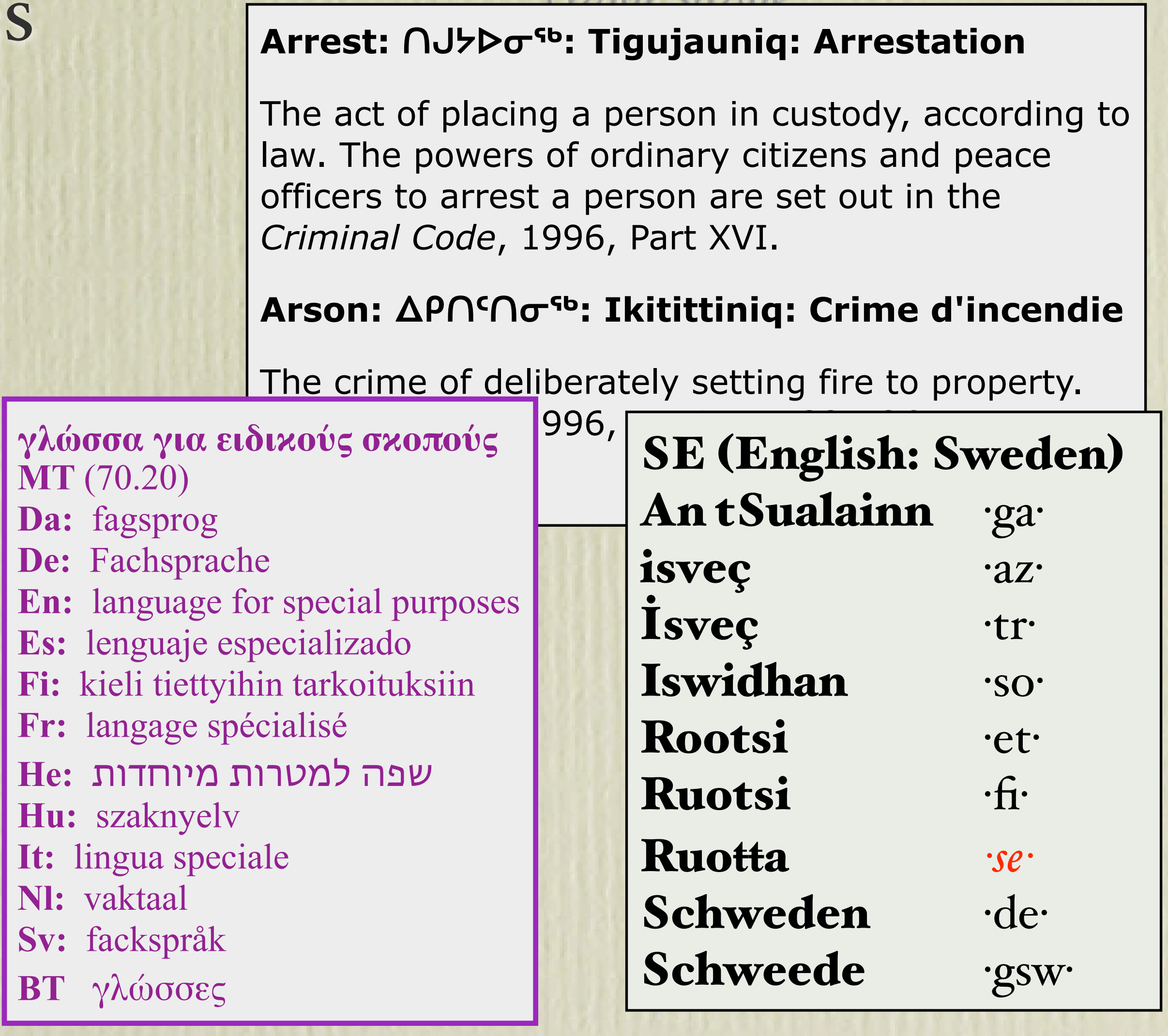
```
                  Table "public.dn"
 Column |  Type   | Modifiers | Storage | Description
--------+---------+-----------+---------+-------------
 dn     | integer | not null  | plain   | ID
 mn     | integer | not null  | plain   | meaning
 ex     | integer | not null  | plain   | expression
Indexes:
    "dn_pkey" PRIMARY KEY, btree (dn)
    "dn_mn_ex_key" UNIQUE, btree (mn, ex) CLUSTER
    "dn_ex_idx" btree (ex)
    "dn_mn_idx" btree (mn)
Foreign-key constraints:
    "dn_ex_fkey" FOREIGN KEY (ex) REFERENCES ex(ex)
    "dn_mn_fkey" FOREIGN KEY (mn) REFERENCES mn(mn)
Referenced by:
    TABLE "md" CONSTRAINT "md_dn_fkey" FOREIGN KEY (dn) REFERENCES dn(dn)
    TABLE "pl0" CONSTRAINT "pl0_mn_fkey" FOREIGN KEY (mn, ex) REFERENCES dn(mn, ex)
        ON UPDATE CASCADE ON DELETE CASCADE
    TABLE "pl1" CONSTRAINT "pl1_mn_fkey1" FOREIGN KEY (mn, ex) REFERENCES dn(mn, ex)
        ON UPDATE CASCADE ON DELETE CASCADE
    TABLE "wc" CONSTRAINT "wc_dn_fkey" FOREIGN KEY (dn) REFERENCES dn(dn)
Triggers:
    dnexap AFTER INSERT OR DELETE OR UPDATE ON dn FOR EACH ROW EXECUTE PROCEDURE exap()
```

Construction

Step 1. Acquire resources.

- Monolingual dictionaries
- Bilingual dictionaries
- Multilingual dictionaries
- Wiktionaries
- Glossaries
- Standards
- Terminologies
- Wordnets
- Thesauri
- Vocabulary databases
- Locale databases

VikiSözlük
Özgür Sözlük



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)

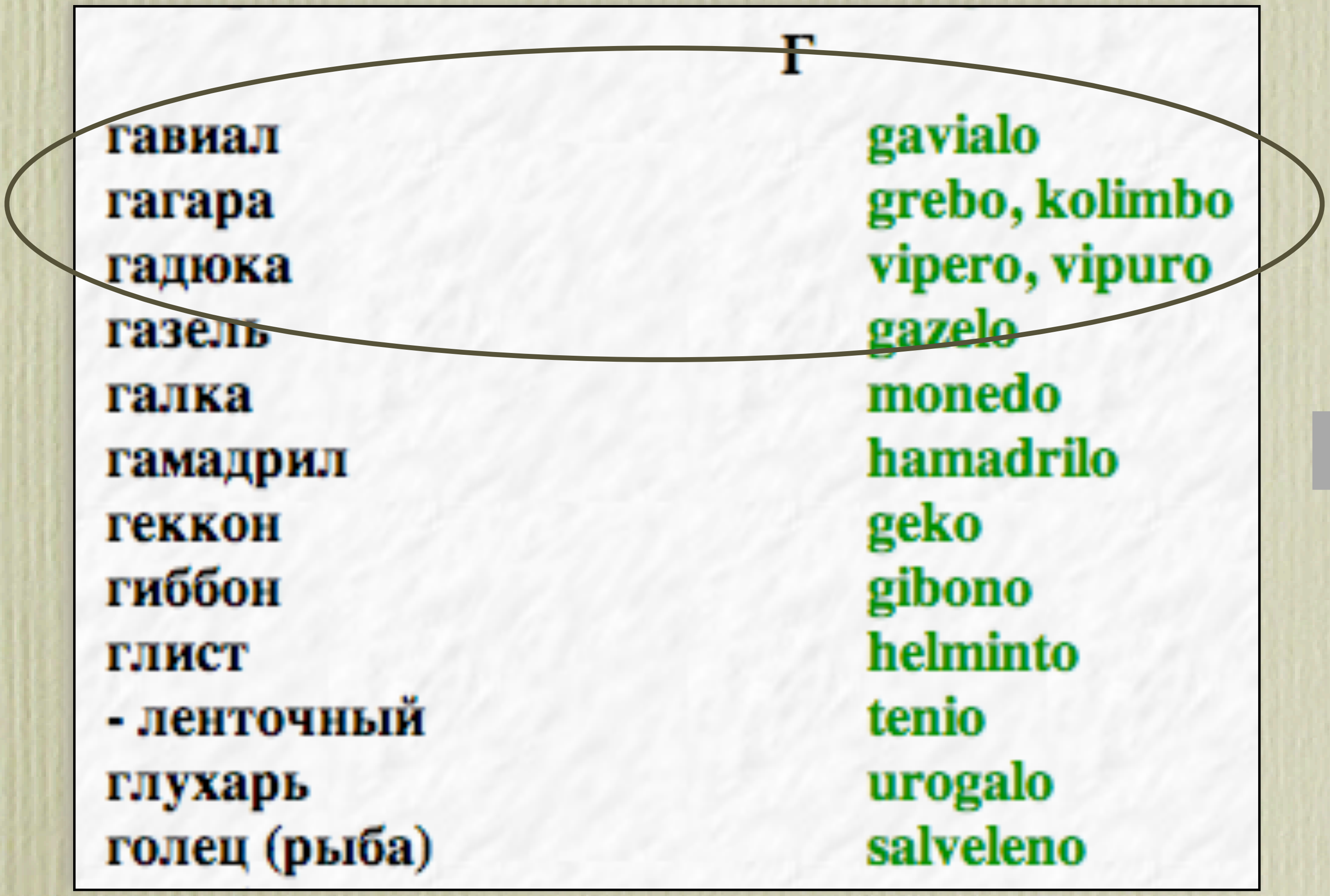
Da:	fagsprog
De:	Fachsprache
En:	language for special purposes
Es:	lenguaje especializado
Fi:	kieli tiettyihin tarkoituksiin
Fr:	langage spécialisé
He:	שפה למטרות מיוחדות
Hu:	szaknyelv
It:	lingua speciale
Nl:	vaktaal
Sv:	fackspråk
BT	γλώσσες

SE (English: Sweden)

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

Construction

Step 2. Obtain and normalize facts from resources.

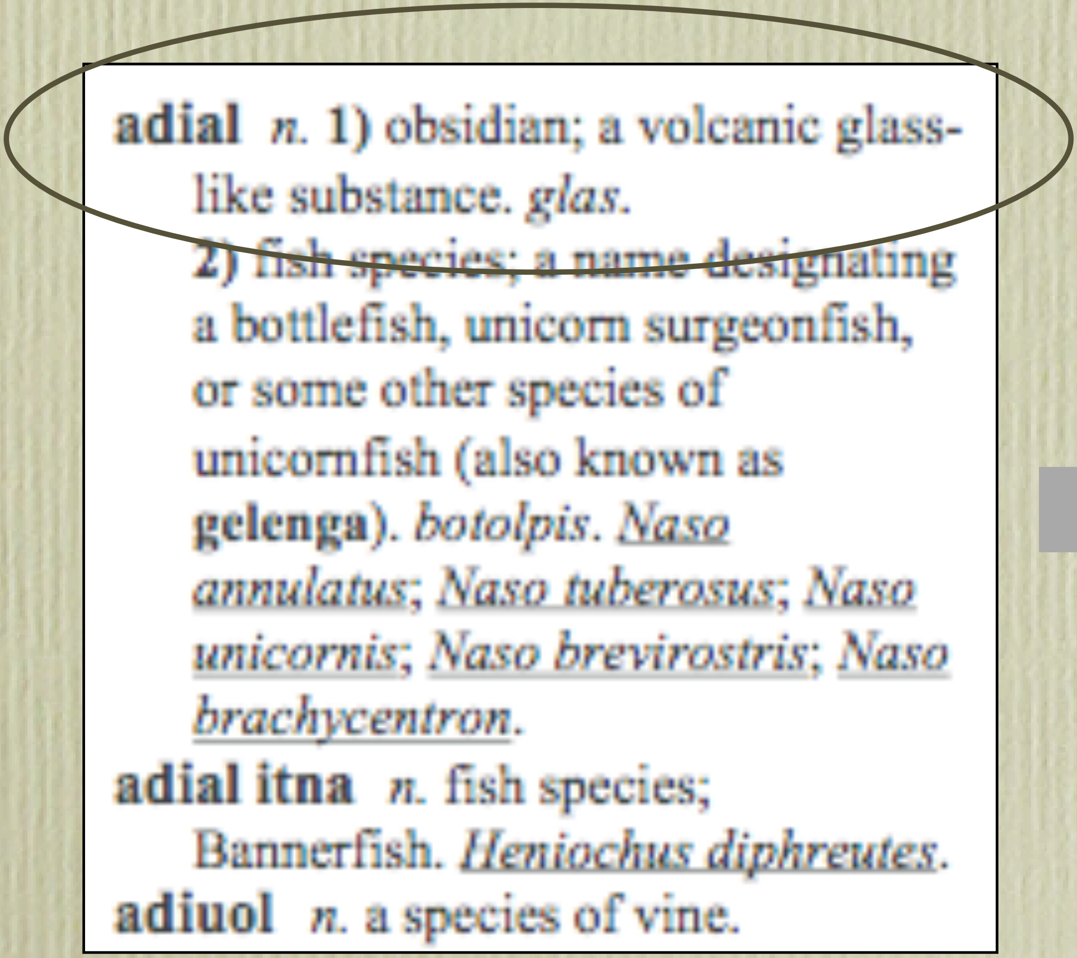


Г	
гавиал	gavialo
гагара	grebo, kolimbo
гадюка	vipero, vipuro
газель	gazelo
галка	monedo
гамадрил	hamadrilo
геккон	geko
гиббон	gibono
глист	helminto
- ленточный	tenio
глухарь	urogalo
голец (рыба)	salveleno

```
:
2
rus-000
epo-000

ex
гавиал
ex
gavialo

ex
гагара
ex
grebo

ex
гагара
ex
kolimbo

ex
гадюка
ex
vipero
ex
vipuro
```

Construction

Step 2. Obtain and normalize facts from resources.

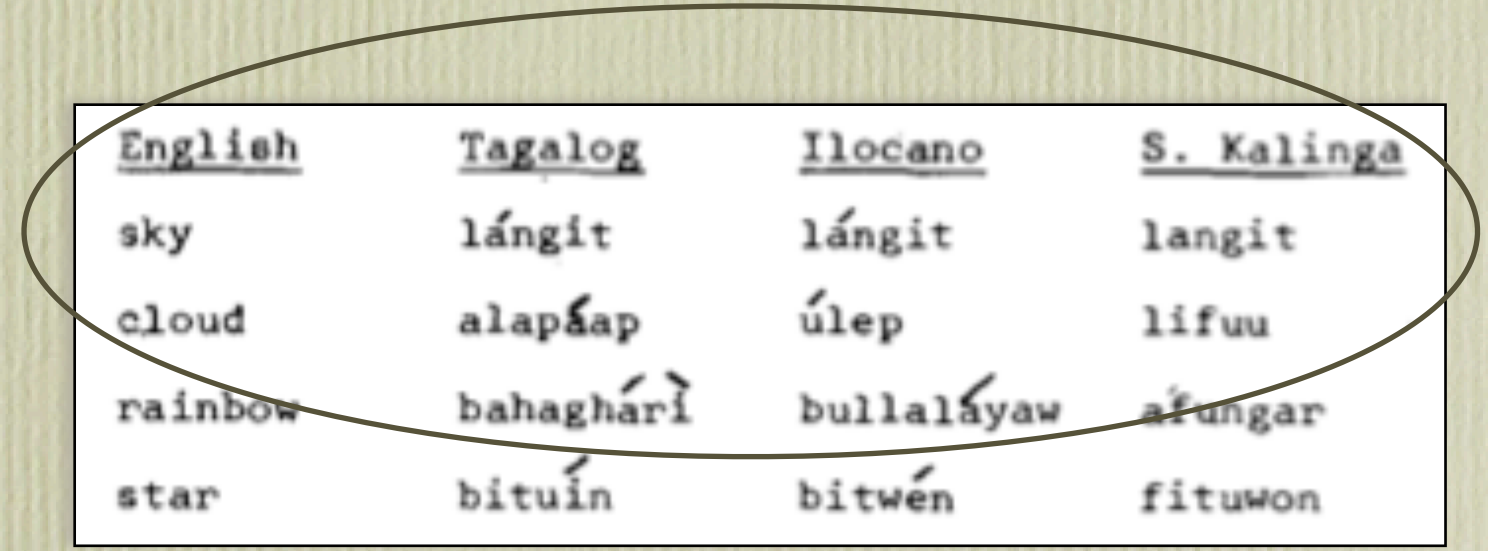


adial *n.* **1)** obsidian; a volcanic glass-like substance. *glas.*
2) fish species; a name designating a bottlefish, unicorn surgeonfish, or some other species of unicornfish (also known as **gelenga**). *botolpis. Naso annulatus; Naso tuberosus; Naso unicornis; Naso brevirostris; Naso brachycentron.*
adial itna *n.* fish species; Bannerfish. *Heniochus diphreutes.*
adiuol *n.* a species of vine.

```
:
1
bch-000

ex
adial
wc
noun
ex
eng-000
obsidian
df
eng-000
a volcanic glass-like substance
ex
tpi-000
glas
```

Construction

Step 2. Obtain and normalize facts from resources.



English	Tagalog	Ilocano	S. Kalinga
sky	lángit	lángit	langit
cloud	alapáap	úlep	lifuu
rainbow	bahaghárì	bullaláyaw	áfungar
star	bitúin	bitwén	fituwon

```
.
1
eng-000

sky
tgl-000
lángit
ilo-000
lángit
ksc-000
langit

cloud
tgl-000
alapáap
ilo-000
úlep
ksc-000
lifuu
```

Construction

Step 2. Obtain and normalize facts from resources.

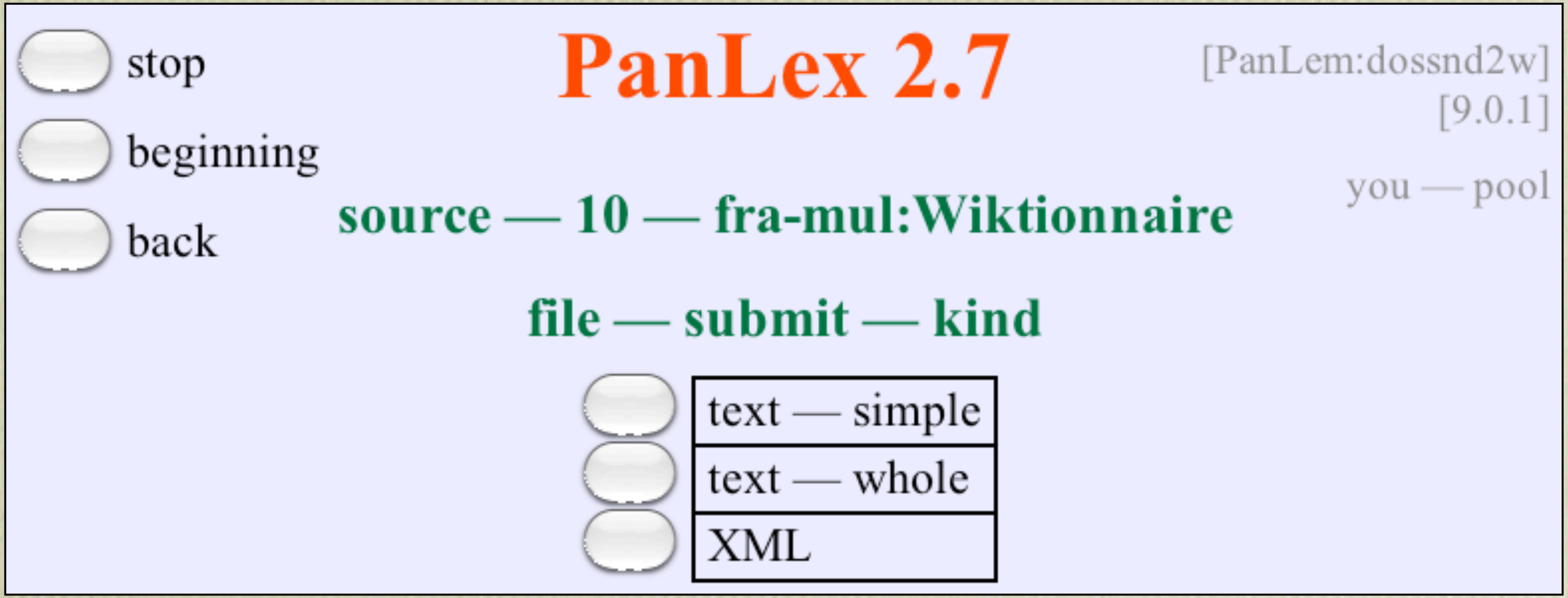
- Character recognition
- Character recoding
- Compositional normalization
- Punctuation standardization
- Lemma standardization
- Word-class standardization
- Entry-structure classification
- Language-variety identification
- Duplicate removal

Jonathan Pool, “Processing LEGO Data for PanLex”, http://www.panlex.org/dev/panlex-lego-prep.html.

Timothy Baldwin, Jonathan Pool, and Susan M. Colowick, “PanLex and LEXTRACT”, Coling 2010.

Construction

Step 3. Add those facts to PanLex.



PanLex 2.7

[PanLem:dossnd2w]
[9.0.1]

you — pool

source — 10 — fra-mul:Wiktionnaire

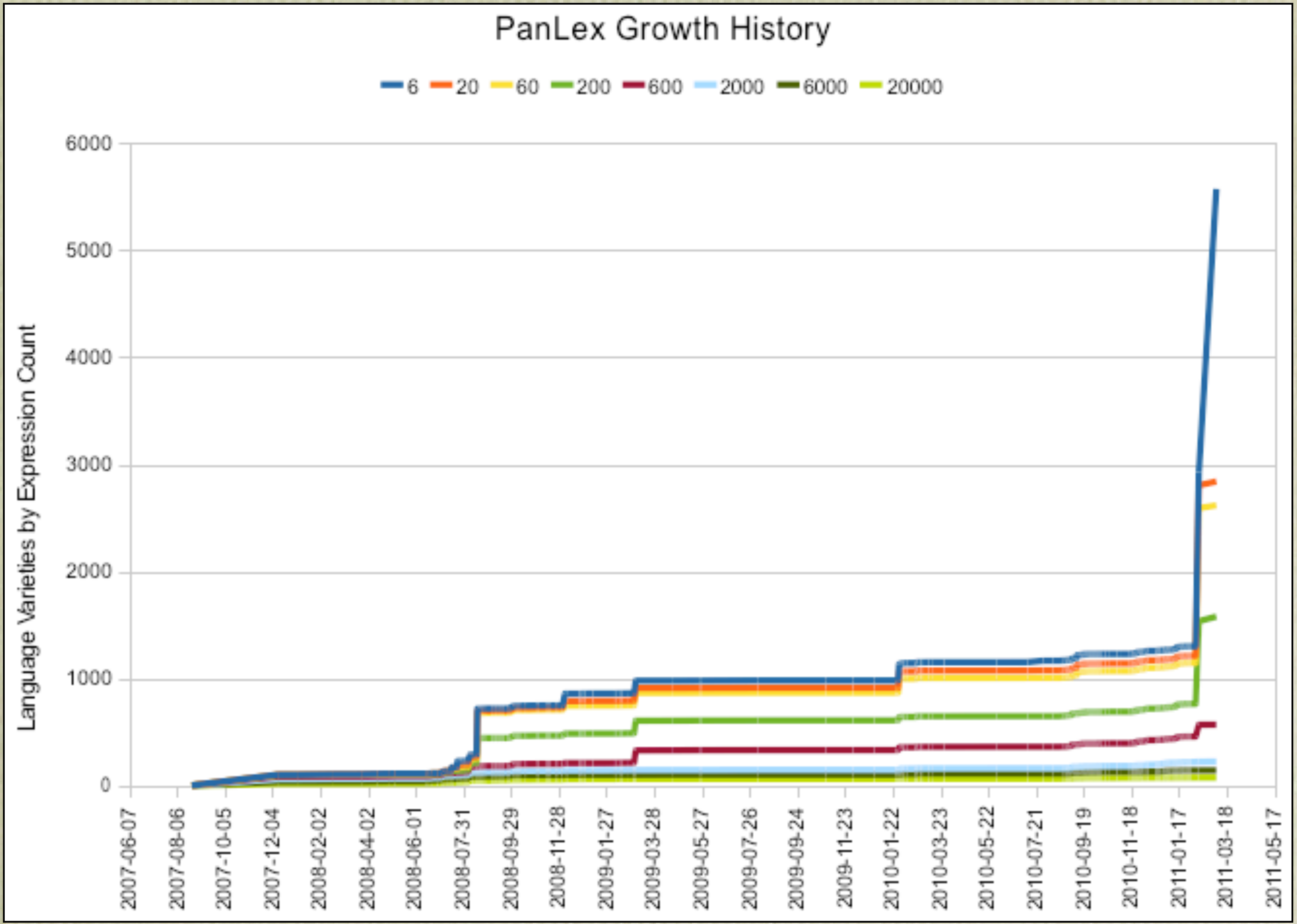
file — submit — kind

Size

Current metrics:

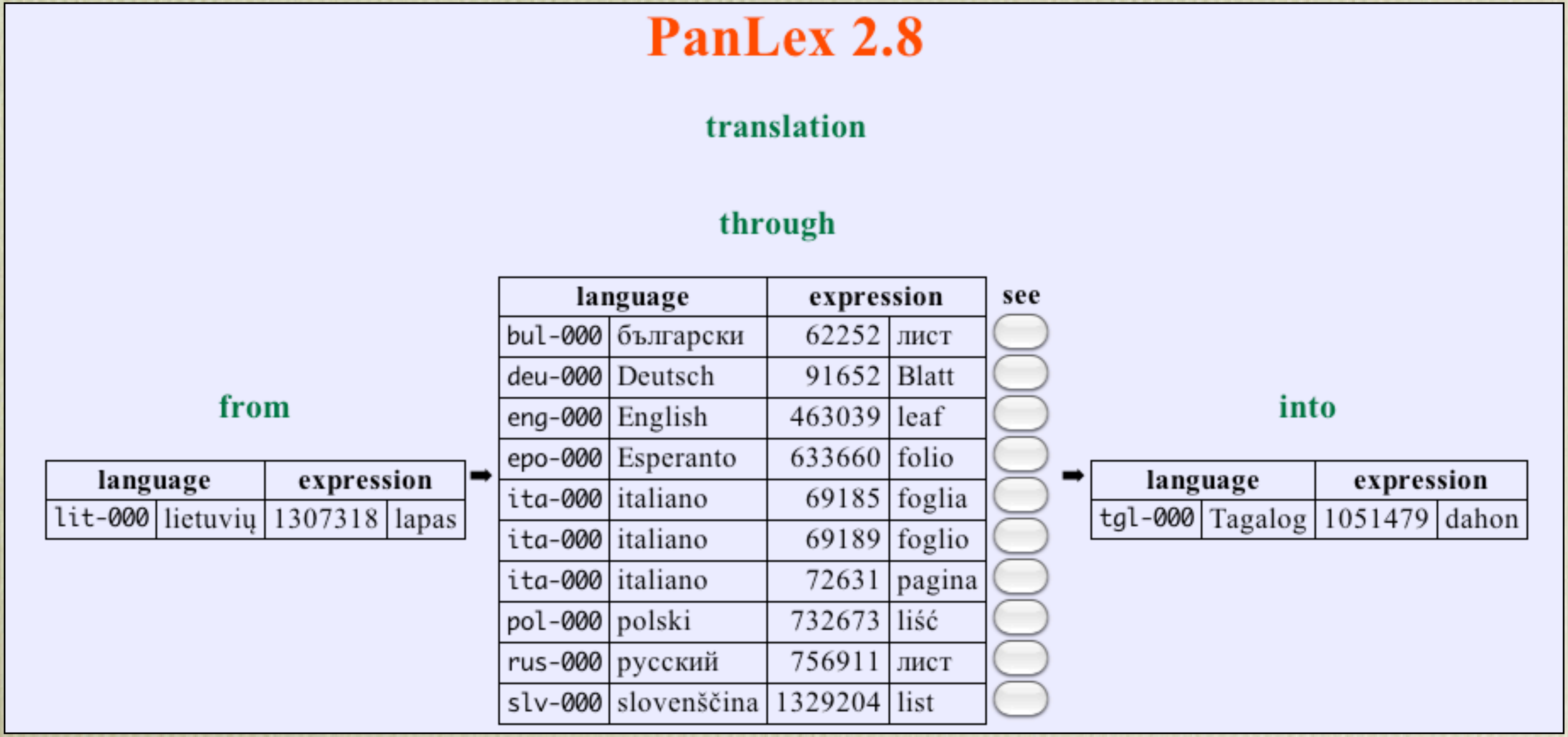
- 17,621,880 expressions
- 6,662 language varieties
- 446,906,568 pairwise translations
- 1,121 sources processed
- 2,202 sources awaiting processing

Size


PanLex Growth History
6
20
60
200
600
2000
6000
20000
Language Varieties by Expression Count
6000
5000
4000
3000
2000
1000
0
2007-06-07
2007-08-06
2007-10-05
2007-12-04
2008-02-02
2008-04-02
2008-06-01
2008-07-31
2008-09-29
2008-11-28
2009-01-27
2009-03-28
2009-05-27
2009-07-26
2009-09-24
2009-11-23
2010-01-22
2010-03-23
2010-05-22
2010-07-21
2010-09-19
2010-11-18
2011-01-17
2011-03-18
2011-05-17

Applications

UIs for PanLex



PanLex 2.8

translation

through

from

language		expression	
lit-000	lietuvių	1307318	lapas

➡

language		expression		see
bul-000	български	62252	лист	○
deu-000	Deutsch	91652	Blatt	○
eng-000	English	463039	leaf	○
epo-000	Esperanto	633660	folio	○
ita-000	italiano	69185	foglia	○
ita-000	italiano	69189	foglio	○
ita-000	italiano	72631	pagina	○
pol-000	polski	732673	liść	○
rus-000	русский	756911	лист	○
slv-000	slovenščina	1329204	list	○

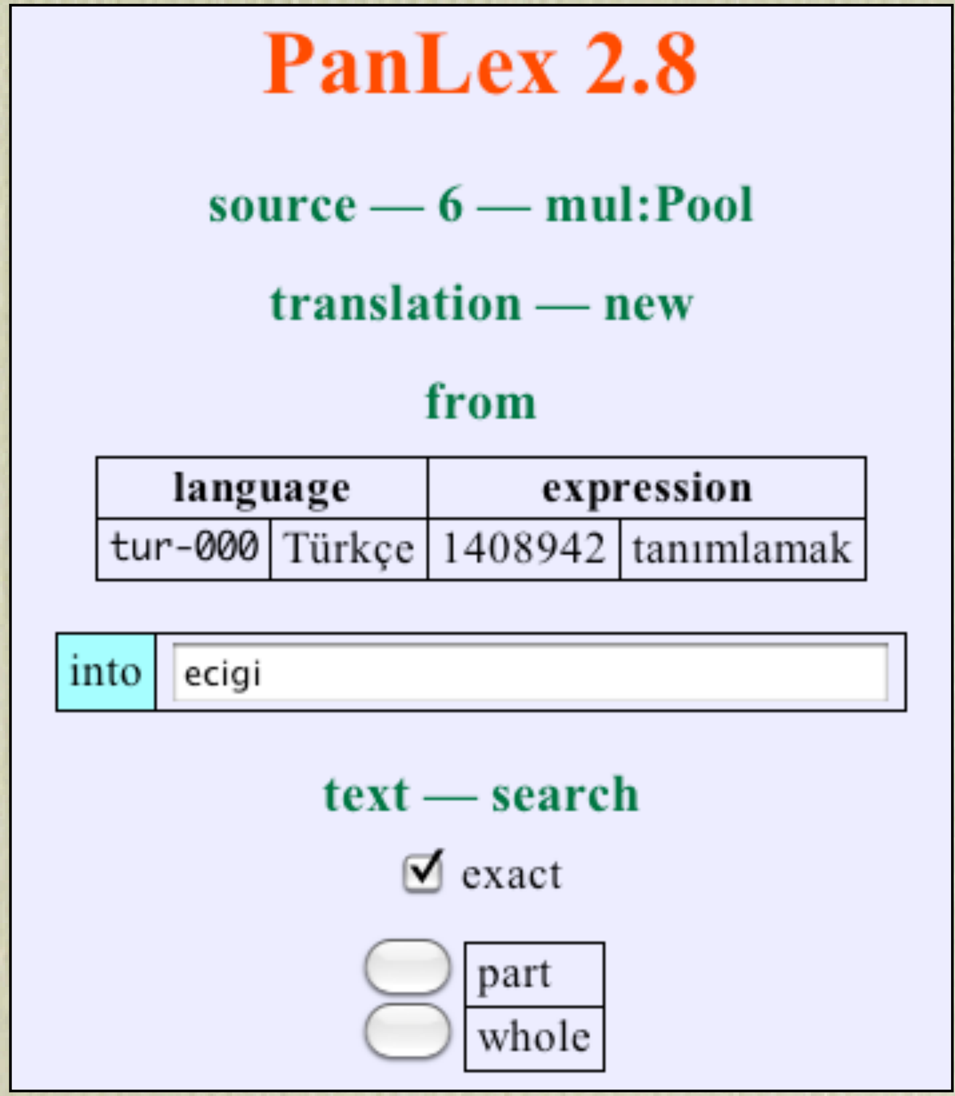
➡

into

language		expression	
tgl-000	Tagalog	1051479	dahon

Applications

UIs for PanLex



PanLex 2.8

source — 6 — mul:Pool

translation — new

from

language		expression	
tur-000	Türkçe	1408942	tanımlamak

into: ecigi

text — search

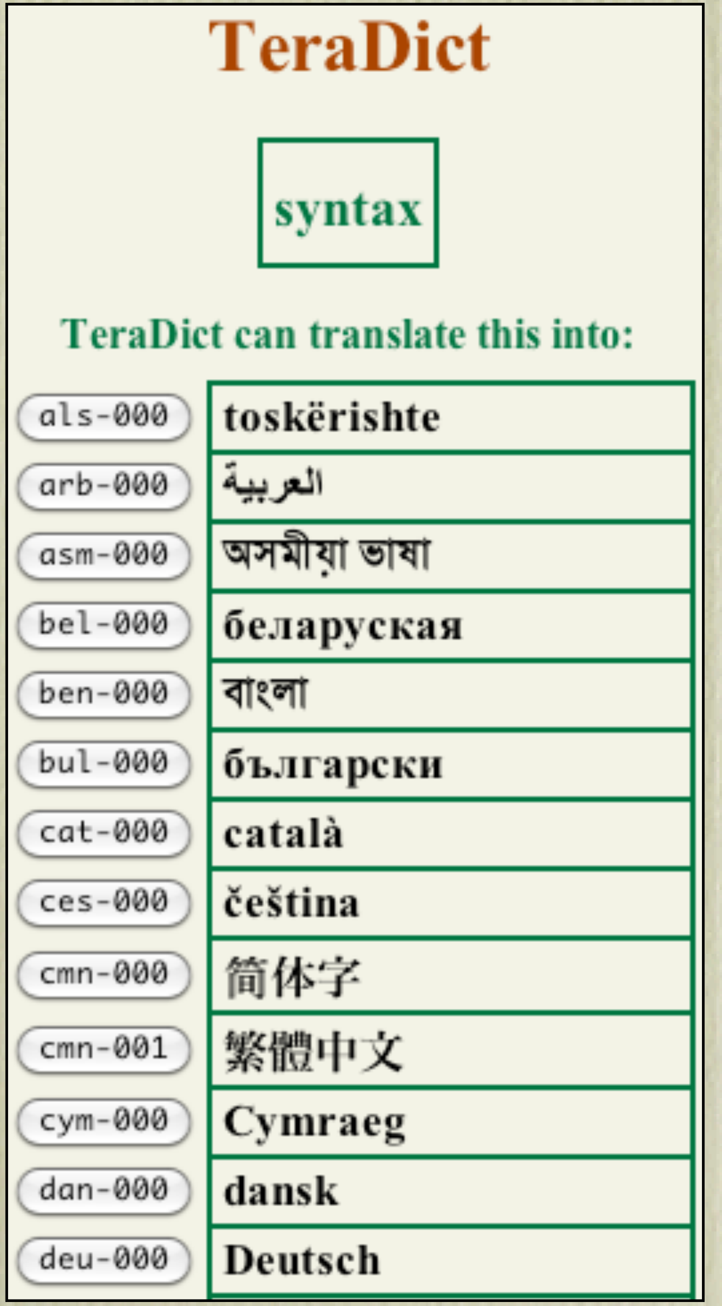
☑ exact

○ part

○ whole

Applications

UIs for PanLex

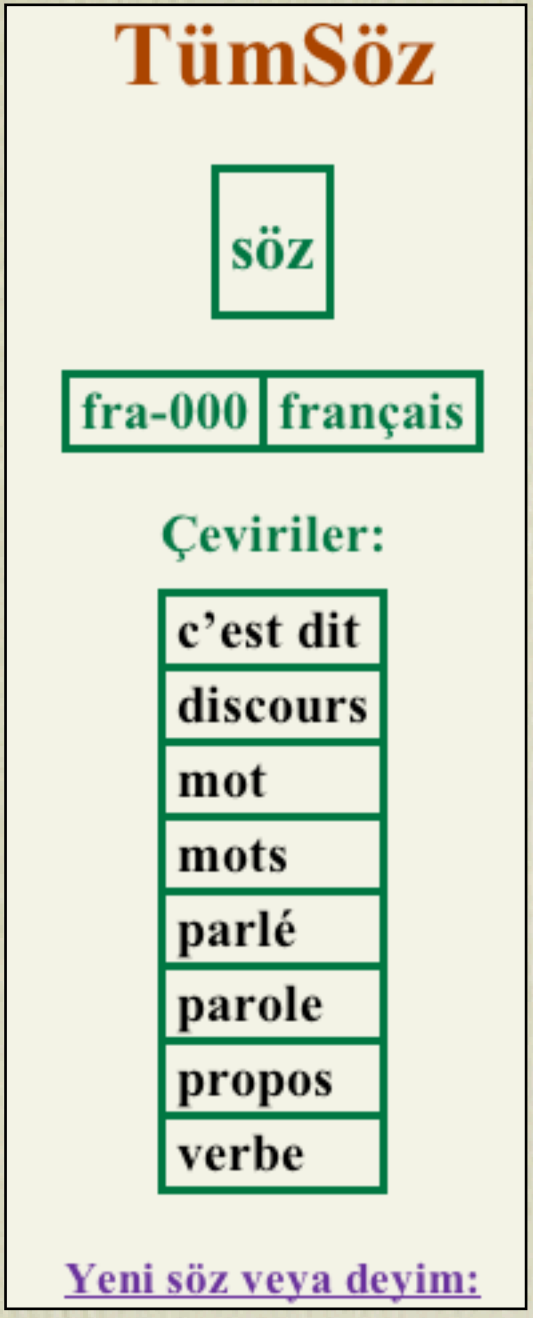


TeraDict

syntax

TeraDict can translate this into:

als-000	toskërishte
arb-000	العربية
asm-000	অসমীয়া ভাষা
bel-000	беларуская
ben-000	বাংলা
bul-000	български
cat-000	català
ces-000	čeština
cmn-000	简体字
cmn-001	繁體中文
cym-000	Cymraeg
dan-000	dansk
deu-000	Deutsch



TümSöz

söz

fra-000 | français

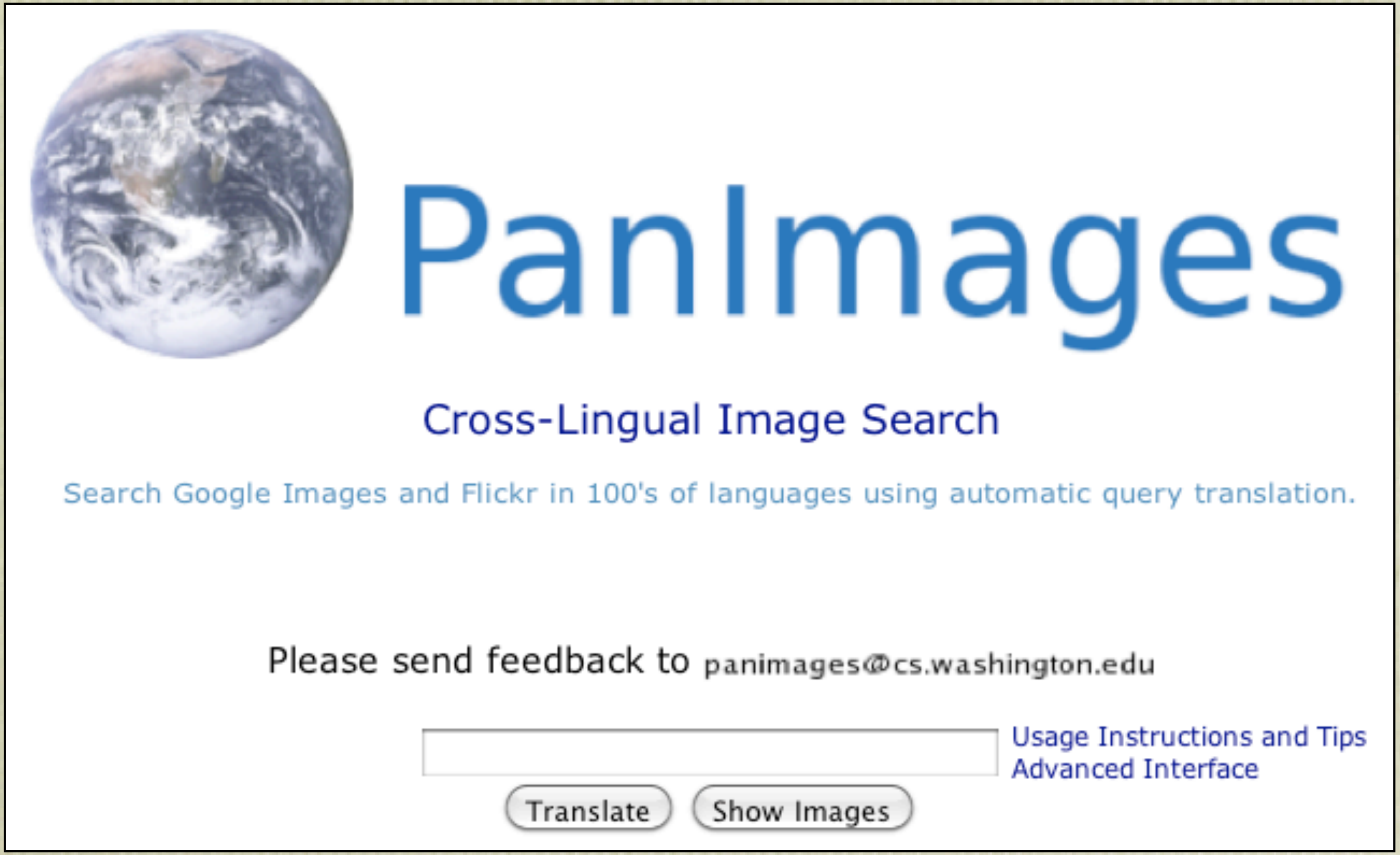
Çeviriler:

- c’est dit
- discours
- mot
- mots
- parlé
- parole
- propos
- verbe

Yeni söz veya deyim:

Applications

Search



PanImages

Cross-Lingual Image Search

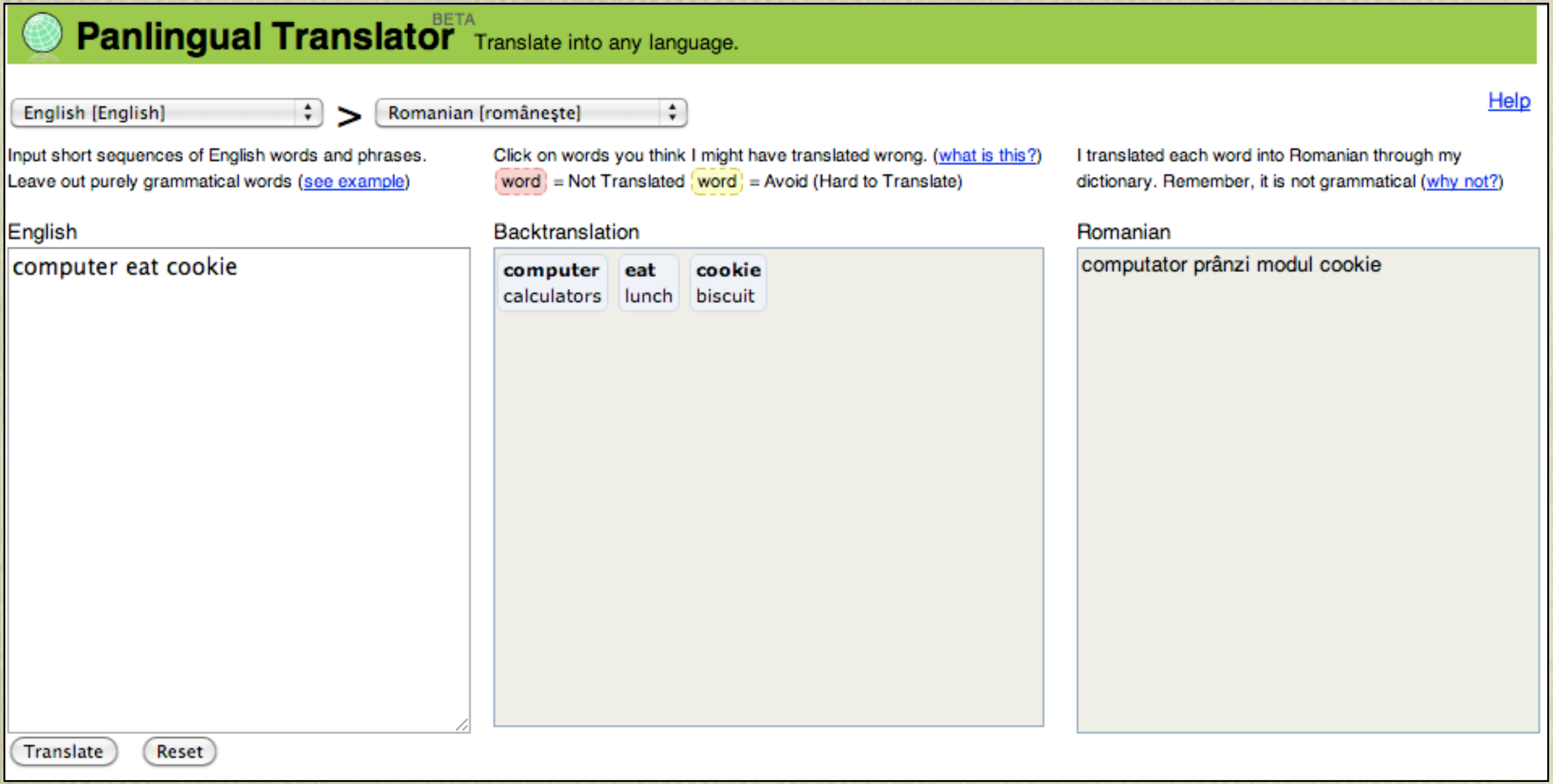
Search Google Images and Flickr in 100's of languages using automatic query translation.

Please send feedback to panimages@cs.washington.edu

Janara Christensen, Mausam, and Oren Etzioni, "A Rose is a Roos is a Ruusu: Querying Translatins for Web Image Search", ACL-IJCNLP 2009.

Applications

Translation



English [English] > Romanian [românește]

Input short sequences of English words and phrases.
Leave out purely grammatical words (see example)

Click on words you think I might have translated wrong. (what is this?)
word = Not Translated word = Avoid (Hard to Translate)

I translated each word into Romanian through my dictionary. Remember, it is not grammatical (why not?)

English

computer eat cookie

Backtranslation

computer	eat	cookie
calculators	lunch	biscuit

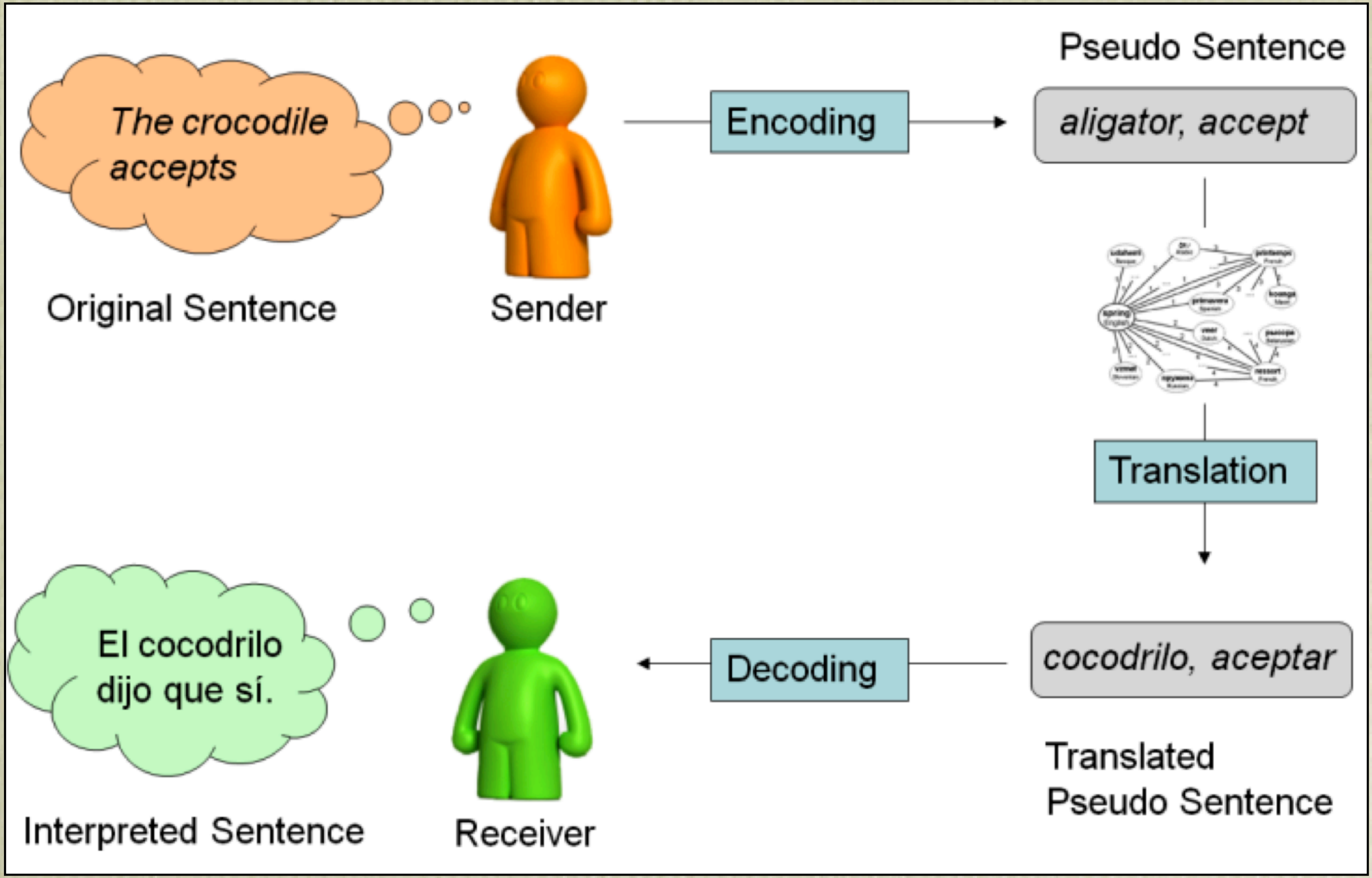
Romanian

computator prânzi modul cookie

Stephen Soderland, Christopher Lim, Mausam, Bo Qin, Oren Etzioni, and Jonathan Pool, "Lemmatic Machine Translation", Proceedings of Machine Translation Summit XII, 2009.

Applications

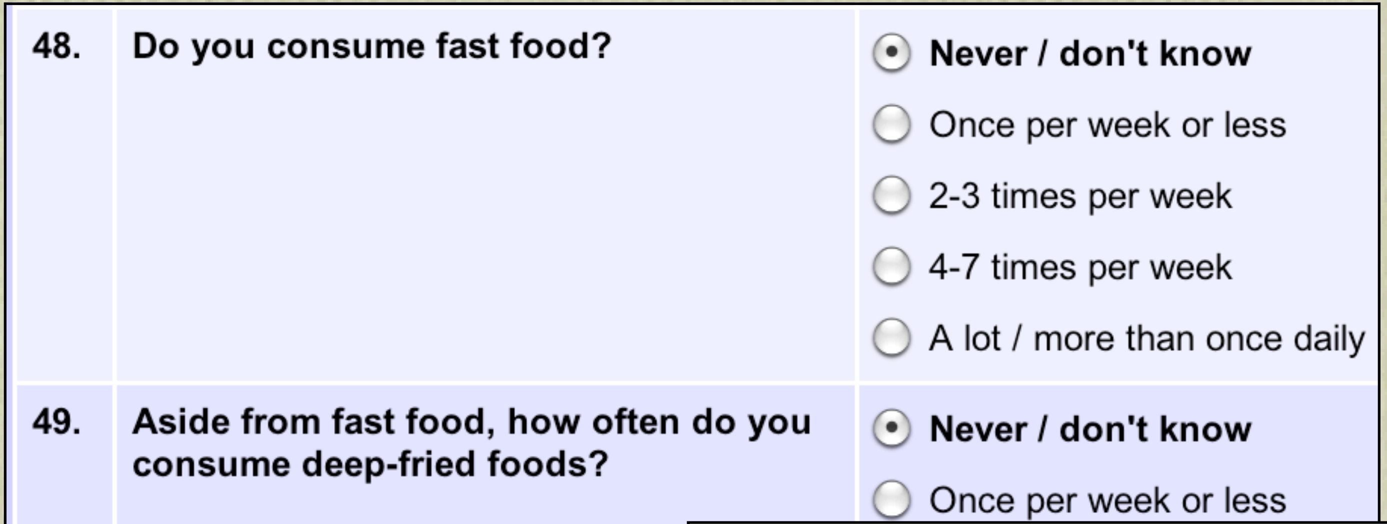
Lemmatic Communication



Katherine Everitt, Christopher Lim, Oren Etzioni, Jonathan Pool, Susan Colowick, Stephen Soderland, "Evaluating Lemmatic Communication", *trans-kom*, 3, 2010, 70–84.

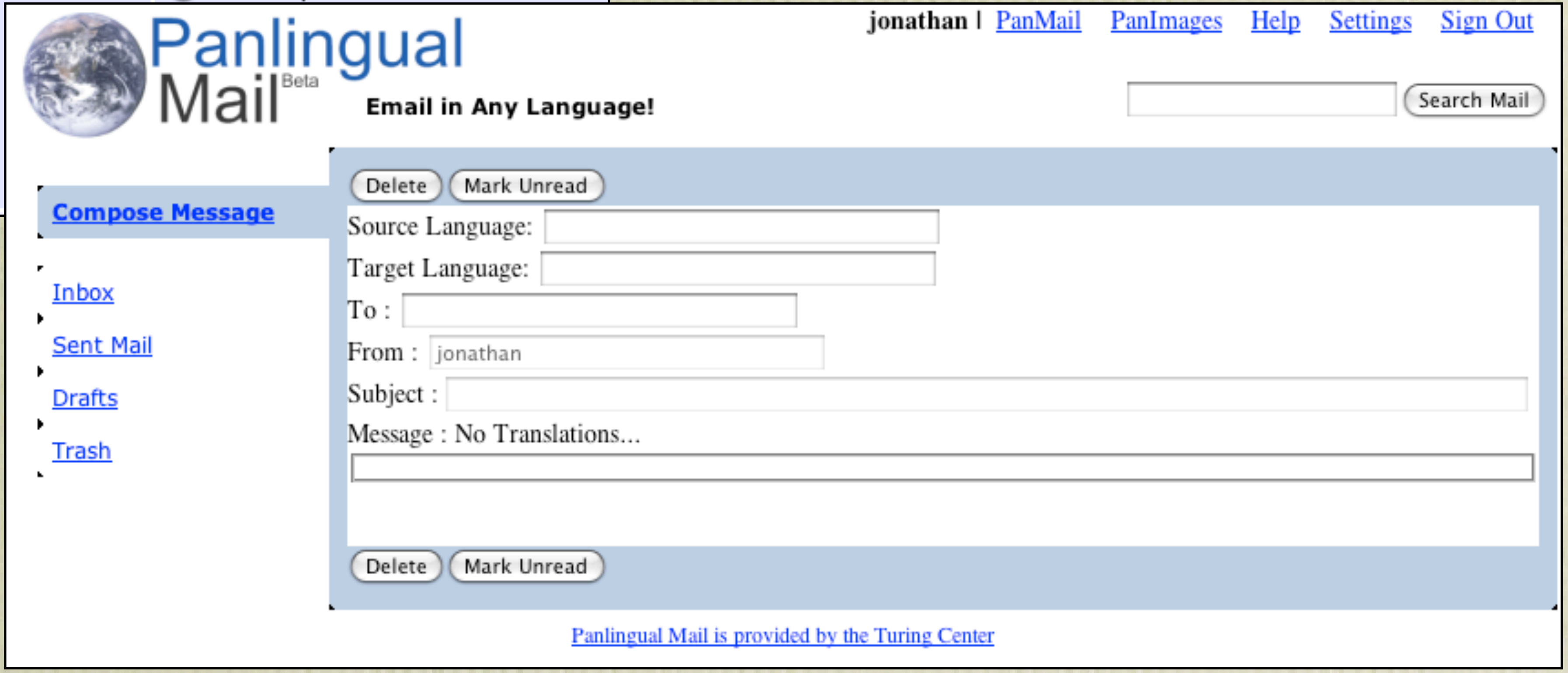
Applications

Polling?



48.	Do you consume fast food?	● **Never / don't know** ○ Once per week or less ○ 2-3 times per week ○ 4-7 times per week ○ A lot / more than once daily
49.	Aside from fast food, how often do you consume deep-fried foods?	● **Never / don't know** ○ Once per week or less

Messaging?



Source Language:

Target Language:

To :

From : jonathan

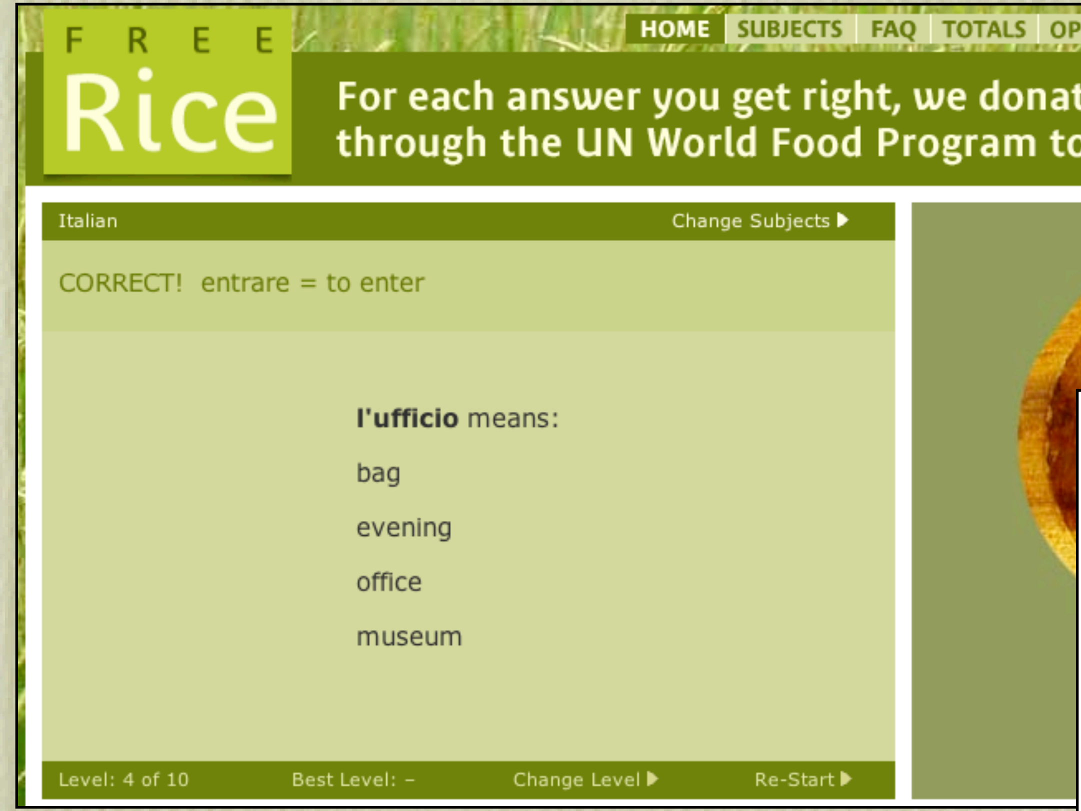
Subject :

Message : No Translations...

Panlingual Mail is provided by the Turing Center

Applications

Games?



For each answer you get right, we donate 10 grains of rice through the UN World Food Program to help end hunger

Italian

CORRECT! entrare = to enter

l'ufficio means:

bag

evening

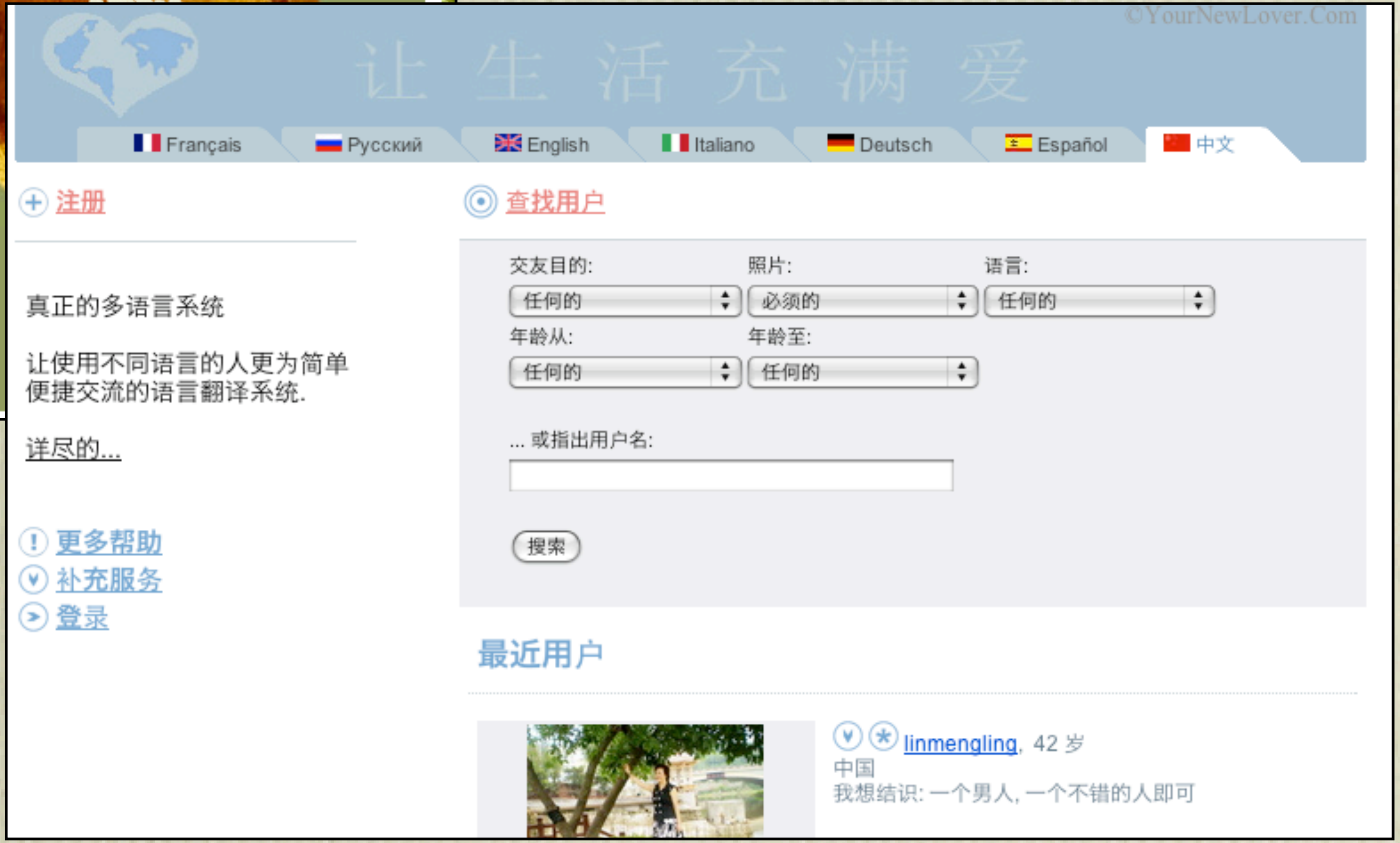
office

museum

Level: 4 of 10

Best Level: –

Social networking?



真正的多语言系统

让使用不同语言的人更为简单便捷交流的语言翻译系统.

详尽的...

交友目的: 任何的

照片: 必须的

语言: 任何的

年龄从: 任何的

年龄至: 任何的

... 或指出用户名:

最近用户

linmengling, 42 岁
中国
我想结识: 一个男人, 一个不错的人即可

Applications

Subtitles?

Current work

- Process 2,000+ acquired sources
- Acquire new digital sources
- Digitize and process printed sources
- Cultivate partnerships and volunteers

http://utilika.org/info/intern2011.html

Current work

Internship projects, 2011–2012

- Lexical acquisition:
  - ✓ Optical character recognition
  - ✓ Lexicographic parsing
  - ✓ Game development

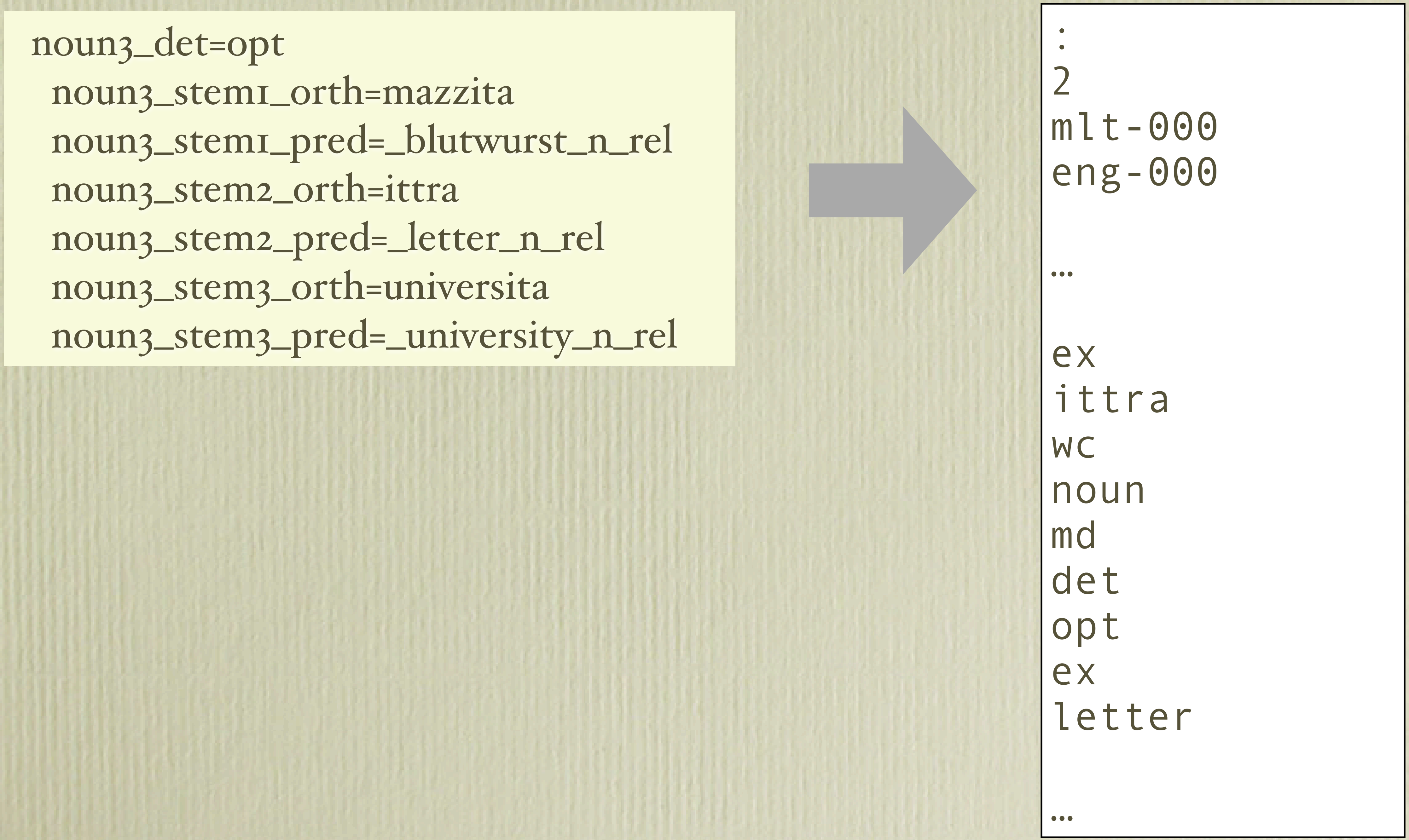
- Applications:
  - ✓ Image search
  - ✓ Mobile app
  - ✓ Social app

- Infrastructure:
  - ✓ Translation inference
  - ✓ Grammar engineering
  - ✓ Graph visualization

http://utilika.org/info/intern2011.html

PanLex and Grammar Engineering

Lexical acquisition **from** grammars

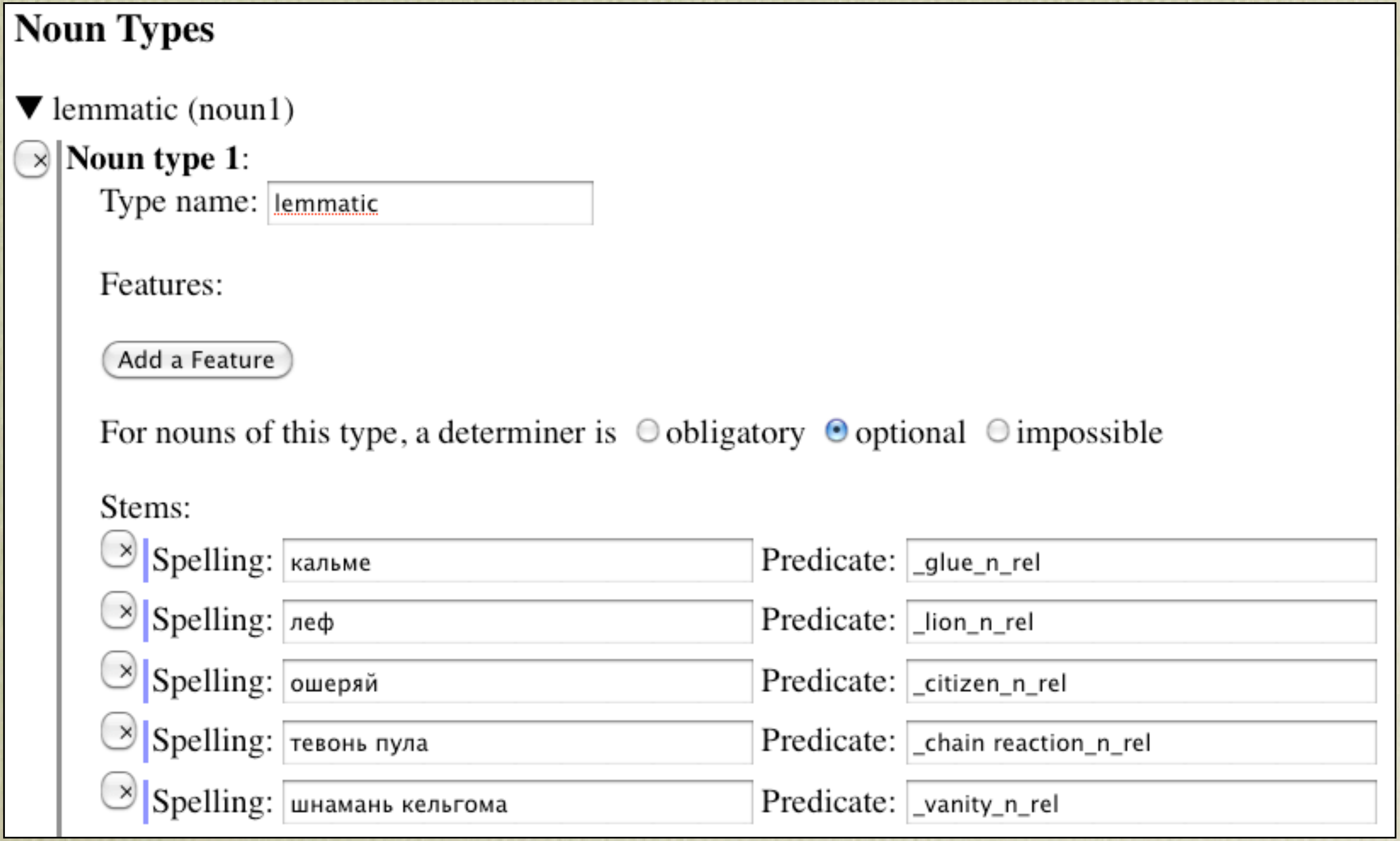


```
noun3_det=opt
  noun3_stem1_orth=mazzita
  noun3_stem1_pred=_blutwurst_n_rel
  noun3_stem2_orth=ittra
  noun3_stem2_pred=_letter_n_rel
  noun3_stem3_orth=universita
  noun3_stem3_pred=_university_n_rel
```

```
:
2
mlt-000
eng-000

…

ex
ittra
wc
noun
md
det
opt
ex
letter

…
```

PanLex and Grammar Engineering

Lexemes and predicates **for** grammars



Noun Types

▼ lemmatic (noun1)

Noun type 1:

Type name: lemmatic

Features:

For nouns of this type, a determiner is ○ obligatory ● optional ○ impossible

Stems:

Spelling:	кальме	Predicate:	_glue_n_rel
Spelling:	леф	Predicate:	_lion_n_rel
Spelling:	ошеряй	Predicate:	_citizen_n_rel
Spelling:	тевонь пула	Predicate:	_chain reaction_n_rel
Spelling:	шнамань кельгома	Predicate:	_vanity_n_rel

PanLex and Grammar Engineering

Instantly generate 6,000 minimal grammars

```
version=6

section=language
language=Bandjalang

section=word-order
word-order=free
has-dets=no
has-aux=no

section=number

section=person
person=none
```

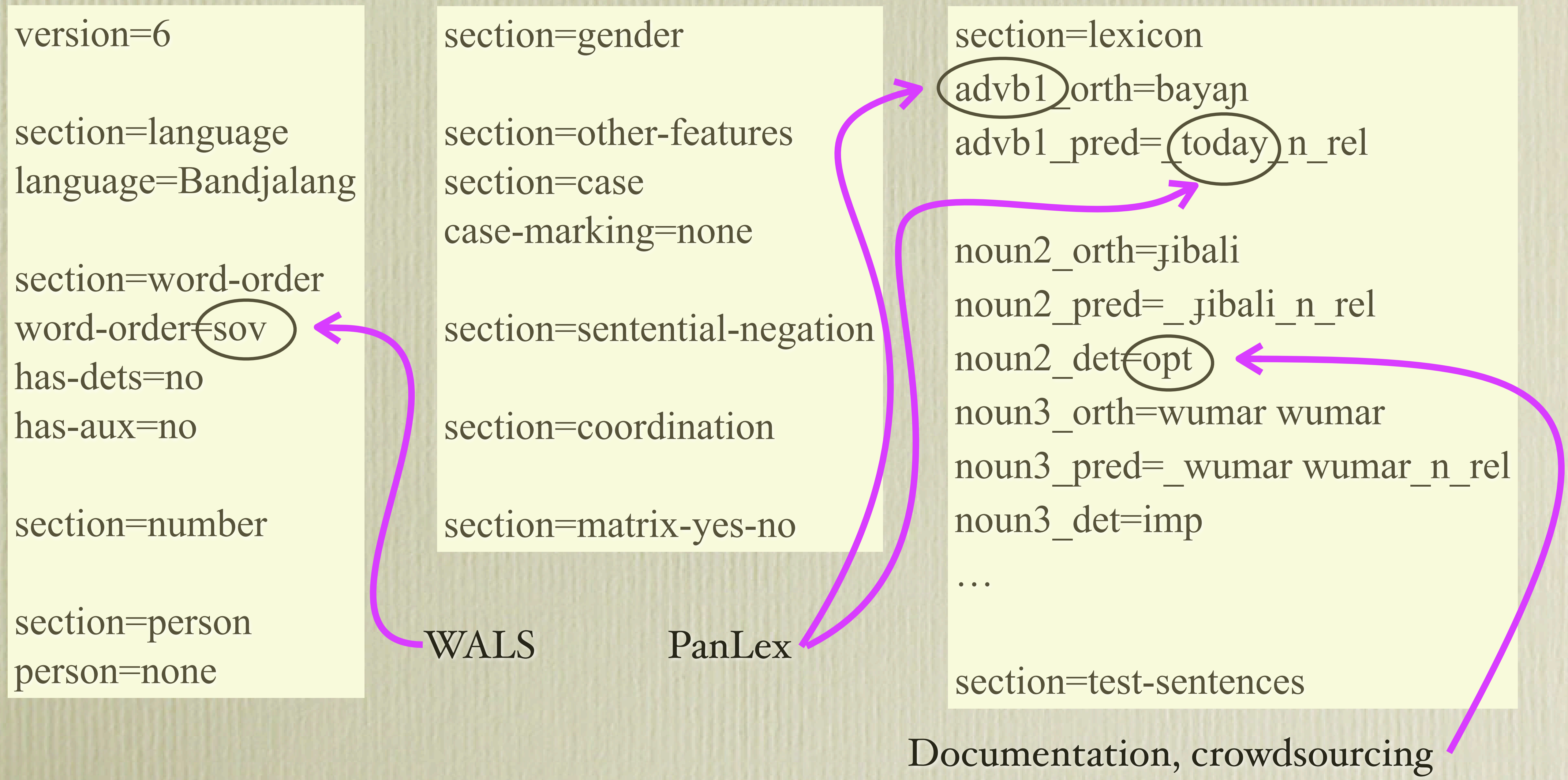
```
section=gender

section=other-features
section=case
case-marking=none

section=sentential-negation

section=coordination

section=matrix-yes-no
```

```
section=lexicon
noun1_orth=bayaɲ
noun1_pred=_bayaɲ_n_rel
noun1_det=imp
noun2_orth=ɟibali
noun2_pred=_ ɟibali_n_rel
noun2_det=imp
noun3_orth=wumar wumar
noun3_pred=_wumar wumar_n_rel
noun3_det=imp
…

section=test-sentences
```

Cf. David Wax,
Matrix ODIN Mash-up

PanLex and Grammar Engineering

Then incrementally differentiate and relate the grammars



```
version=6

section=language
language=Bandjalang

section=word-order
word-order=sov
has-dets=no
has-aux=no

section=number

section=person
person=none
```

```
section=gender

section=other-features
section=case
case-marking=none

section=sentential-negation

section=coordination

section=matrix-yes-no
```

```
section=lexicon
advb1_orth=bayaɲ
advb1_pred=_today_n_rel

noun2_orth=ɟibali
noun2_pred=_ɟibali_n_rel
noun2_det=opt
noun3_orth=wumar wumar
noun3_pred=_wumar wumar_n_rel
noun3_det=imp
…

section=test-sentences
```

WALS

PanLex

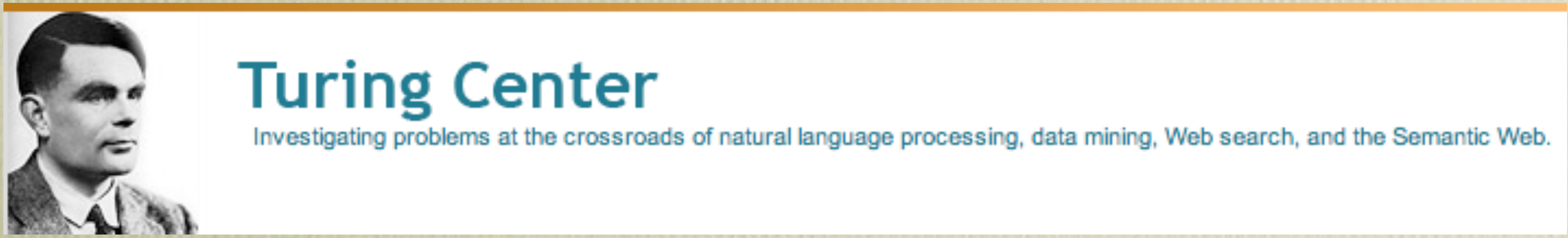
Documentation, crowdsourcing

Team

Turing Center, University of Washington

http://www.turing.washington.edu

- Oren Etzioni
- Katherine Everett
- Christopher Lim
- Mausam
- Kobi Reiter
- Marcus Sammer
- Michael Schmitz
- Michael Skinner
- Stephen Soderland



Turing Center

Investigating problems at the crossroads of natural language processing, data mining, Web search, and the Semantic Web.

Team

Utilika Foundation

http://utilika.org

- Susan Colowick
- Jonathan Pool
- Miranda Taylor

Team

Collaborators

- Academic
  - LEGO Project (Timothy Usher, Jeff Good, Helen Aristar-Dry)
  - Timothy Baldwin, University of Melbourne
  - Larry Hyman, UC Berkeley
  - Interns from Univ. of Washington, Univ. of Illinois, Univ. of Maryland, UC Davis, Univ. of Melbourne, CMU, and Ohio State
- Nonprofit
  - Long Now Foundation / Rosetta Project (Laura Welcher)
- For-profit
  - CJK Dictionary Institute (Jack Halpern)
  - Digital Sonata (Vadim Berman)
  - PostgreSQL Experts (Josh Berkus, Quinn Weaver)

Comments/Questions

Info:

http://panlex.org