Automating Dictionary Production: a Tagalog-English-Korean Dictionary from Scratch

Vít Baisa, Marek Blahuš, Michal Cukr, Ondřej Herman, **Miloš Jakubíček**, Vojtěch Kovář, Marek Medveď, Michal Měchura, Pavel Rychlý, Vít Suchomel

LEXiCAL COMPUTING

MASARYK UNIVERSITY

Brno, Czech Republic

eLex 2019, Sintra, Portugal

Automating dictionary production: How far can we get?

data + tools + people = dictionary

About the dictionary

Tagalog → English + Korean

About the dictionary

Tagalog → English + Korean

from scratch, automatically generated draft, post-editing

About the dictionary

Tagalog → English + Korean

from scratch, automatically generated draft, post-editing

15 + 35 = 50

About the dictionary

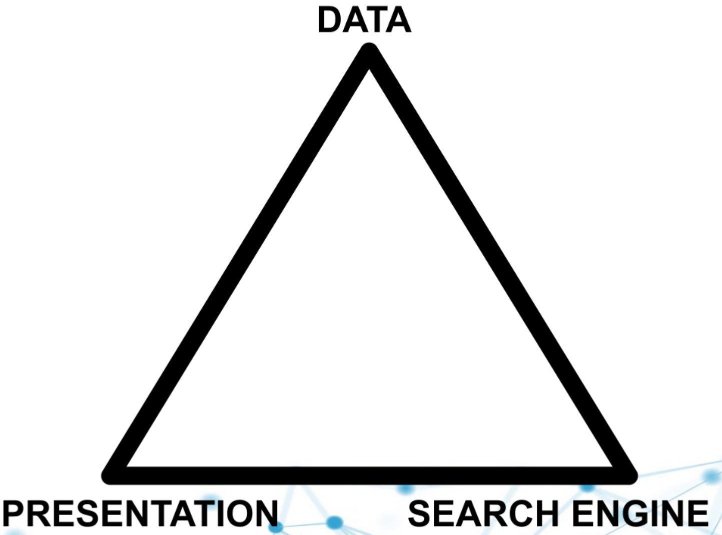
Tagalog → English + Korean

from scratch, automatically generated draft, post-editing

15 + 35 = 50

< 9

Digital only: dictionary is a triangle



Commercial context

Naver corporation, the biggest search engine in South Korea

freely available at https://dict.naver.com/, soon™

Urdu and Lao in the pipeline

Entry structure

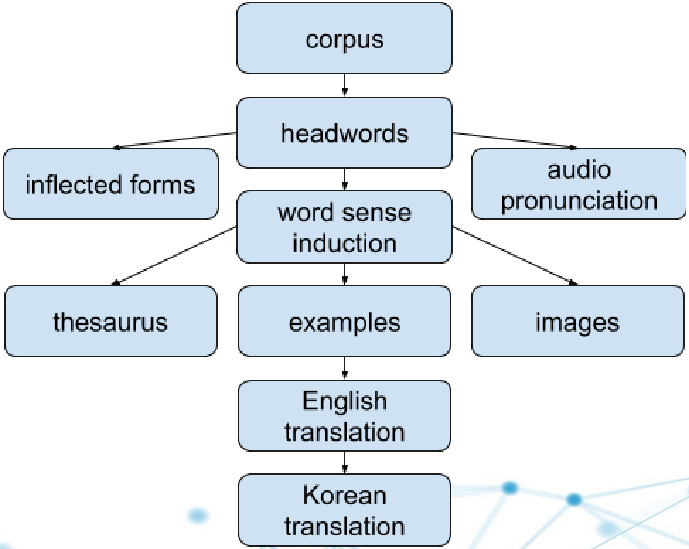
- headword
- inflected forms
- audio pronunciation
- word sense
  - identified with a disambiguating gloss
  - up to 10 collocations
  - up to 10 synonyms/antonyms
  - a picture, if appropriate
  - 3 example sentences
  - English translation of the gloss and one example
  - Korean translation of the gloss and one example

The recipe

- a big web corpus: 650 - 420 = 230
- PoS tagger
- lemmatizer
- sketch grammar
- Sketch Engine
- Lexonomy + custom editing widgets
- 1–8 developers
- 5–15 editors

cook but not boil

Editorial workflow



Editing tasks

small, single, focused tasks

training, multiple annotations, IAA

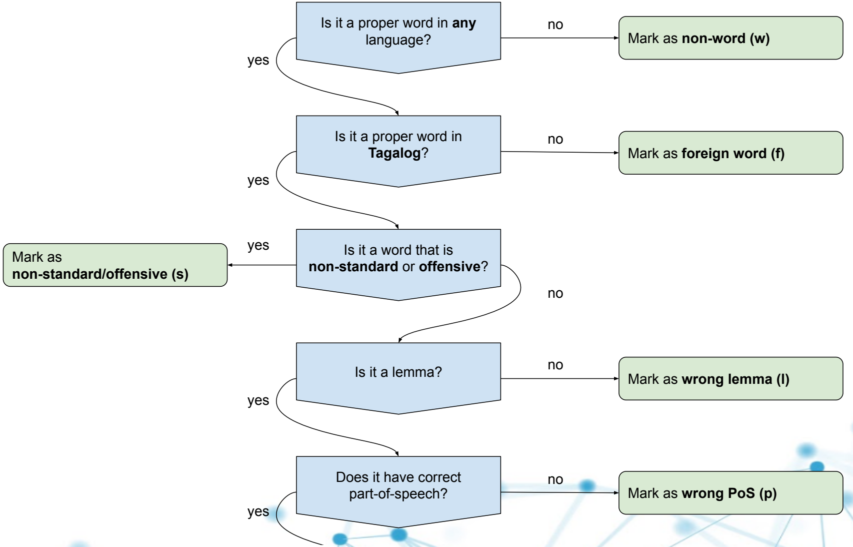
tasks ⇒ batches in Lexonomy

central database in NVH (name-value hierarchy), Měchura (2020)

Corpus

- 650 - 420 = 230
- modified Stanford PoS tagger
- improved stemmer
- sketch grammar

Headwords



Is it a proper word in **any** language?

no

Mark as **non-word (w)**

yes

Is it a proper word in **Tagalog**?

no

Mark as **foreign word (f)**

yes

Is it a word that is **non-standard** or **offensive**?

yes

Mark as **non-standard/offensive (s)**

no

Is it a lemma?

no

Mark as **wrong lemma (l)**

yes

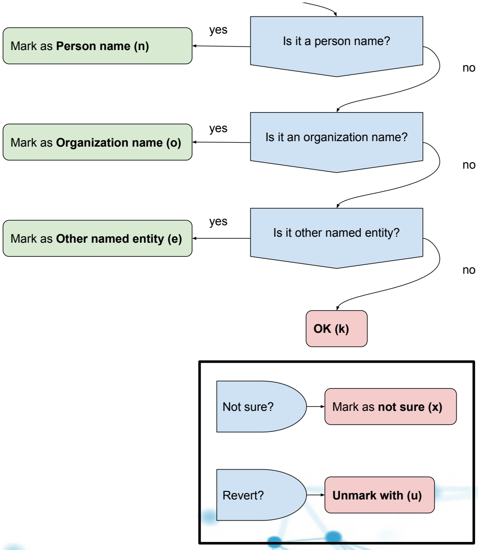
Does it have correct part-of-speech?

no

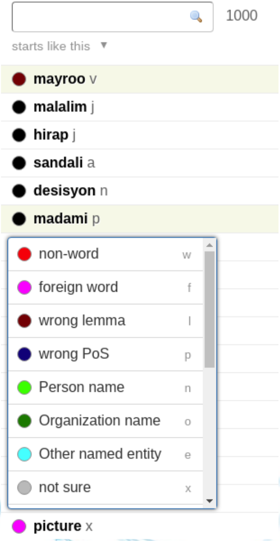
Mark as **wrong PoS (p)**

yes

Headwords


Is it a person name?
yes
Mark as **Person name (n)**
no
Is it an organization name?
yes
Mark as **Organization name (o)**
no
Is it other named entity?
yes
Mark as **Other named entity (e)**
no
OK (k)
Not sure?
Mark as **not sure (x)**
Revert?
Unmark with (u)

Headwords



1000

starts like this ▼

- ● **mayroo** v
- ● **malalim** j
- ● **hirap** j
- ● **sandali** a
- ● **desisyon** n
- ● **madami** p

●	non-word	w
●	foreign word	f
●	wrong lemma	l
●	wrong PoS	p
●	Person name	n
●	Organization name	o
●	Other named entity	e
●	not sure	x

- ● **picture** x

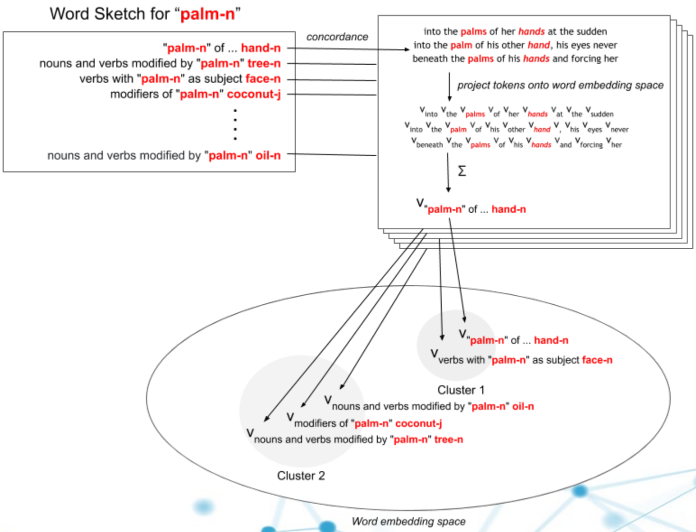
Inflected forms

reassigning lemmas

Pronunciation

Pronunciation

Word senses



Word Sketch for "palm-n"

"palm-n" of ... hand-n

nouns and verbs modified by "palm-n" tree-n

verbs with "palm-n" as subject face-n

modifiers of "palm-n" coconut-j

⋮

nouns and verbs modified by "palm-n" oil-n

concordance

into the palms of her *hands* at the sudden

into the palm of his other *hand*, his eyes never

beneath the palms of his *hands* and forcing her

project tokens onto word embedding space

v_{into} v_{the} v_{palms} v_{of} v_{her} v_{hands} v_{at} v_{the} v_{sudden}

v_{into} v_{the} v_{palm} v_{of} v_{his} v_{other} v_{hand} $v_{,}$ v_{his} v_{eyes} v_{never}

$v_{beneath}$ v_{the} v_{palms} v_{of} v_{his} v_{hands} v_{and} $v_{forcing}$ v_{her}

Σ

$v_{\text{"palm-n" of ... hand-n}}$

$v_{\text{"palm-n" of ... hand-n}}$

$v_{\text{verbs with "palm-n" as subject face-n}}$

Cluster 1

$v_{\text{nouns and verbs modified by "palm-n" oil-n}}$

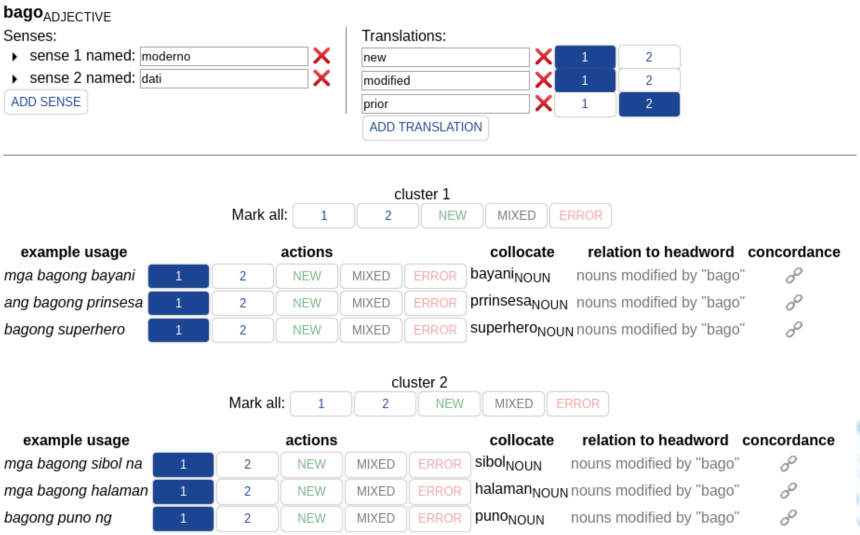
$v_{\text{modifiers of "palm-n" coconut-j}}$

$v_{\text{nouns and verbs modified by "palm-n" tree-n}}$

Cluster 2

Word embedding space

Word senses



bago$_{\text{ADJECTIVE}}$

Senses:

- sense 1 named: moderno ❌
- sense 2 named: dati ❌

ADD SENSE

Translations:

new	❌	[✓ 1]	2
modified	❌	[✓ 1]	2
prior	❌	1	[✓ 2]

ADD TRANSLATION

cluster 1

Mark all: 1 2 NEW MIXED ERROR

example usage	actions	collocate	relation to headword	concordance
mga bagong bayani	[✓ 1] 2 NEW MIXED ERROR	$\text{bayani}_{\text{NOUN}}$	nouns modified by "bago"	🔗
ang bagong prinsesa	[✓ 1] 2 NEW MIXED ERROR	$\text{prrinsesa}_{\text{NOUN}}$	nouns modified by "bago"	🔗
bagong superhero	[✓ 1] 2 NEW MIXED ERROR	$\text{superhero}_{\text{NOUN}}$	nouns modified by "bago"	🔗

cluster 2

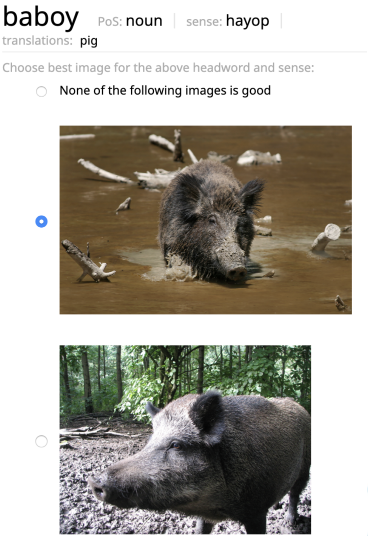
Mark all: 1 2 NEW MIXED ERROR

example usage	actions	collocate	relation to headword	concordance
mga bagong sibol na	[✓ 1] 2 NEW MIXED ERROR	$\text{sibol}_{\text{NOUN}}$	nouns modified by "bago"	🔗
mga bagong halaman	[✓ 1] 2 NEW MIXED ERROR	$\text{halaman}_{\text{NOUN}}$	nouns modified by "bago"	🔗
bagong puno ng	[✓ 1] 2 NEW MIXED ERROR	$\text{puno}_{\text{NOUN}}$	nouns modified by "bago"	🔗

Pictures

- Wikidata + Wiktionary + Wikipedia (“Wikimedia projects”)
- Pixabay
- Google Custom Search

Pictures



baboy PoS: noun | sense: hayop |

translations: pig

Choose best image for the above headword and sense:

○ None of the following images is good

●

○

Collocations

clustering and re-clustering

Thesaurus

from Sketch Engine, on sense level

Examples

GDEX

Translations

GT + MB

Lessons learned

- biggest issues not in technology but in human resource (crowd) and data management
- technologies mature enough to be beneficial
- good and big corpus is the best start for everything
- the better and bigger, the easiest the rest

Conclusions

- How far can we get?
- almost all steps can be efficiently automated and then post-edited
  - if it pays off
- Tagalog almost finished and soon to be published
- Urdu, Lao in the pipeline
- more detailed information on post-editing statistics to be published