european lexicographic infrastructure

D4.7 Evaluation and assessment of methods for automatic drafting

Authors: Miloš Jakubíček, Vojtěch Kovář, Adam Rambousek

Date: July 31st, 2022

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D4.7 Evaluation and assessment of methods for automatic enriching of lexicographic resources

Deliverable Number: D4.7
Dissemination Level: Public
Delivery Date: July 31, 2022
Version: 1
Authors: Miloš Jakubíček, Vojtěch Kovář, Adam Rambousek, Marek Blahuš

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
1, July 31st	Submission-ready	Miloš Jakubíček, Vojtěch Kovář, Adam Rambousek

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Introduction

This document contains evaluation of the methods for automatic drafting of dictionaries that were subject to the D4.1 Online Dictionary Post-Editing and Presentation Module and D4.2 Dictionary Drafting Module. We describe the experiments performed, mostly in the production environment, and present an overall assessment of the related methodology.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

1. Background: dictionary post-editing

The relationship between lexicography and text corpora has been well described in [2] in terms of “corpus revolutions”.

The first corpus revolution was when the corpus was born as a digital medium representing the source of empirical evidence in linguistics and in lexicography in particular so that linguistic introspection could be largely replaced by language evidence.

The second corpus revolution happened when the size of the corpora started growing. On one hand, this allowed lexicographers to get more reliable evidence for more words and multi-word expressions, on the other hand it was no longer feasible to inspect corpus contents manually by mere concordances. Sophisticated extraction tools like Sketch Engine [1] had to be developed so that lexicographers could analyze multi-billion corpora efficiently.

This deliverable addresses the third corpus revolution that is happening now: the post-editing revolution. Using advanced natural language processing tools and methods it is possible to construct a whole dictionary draft fully automatically and let lexicographers only correct, i.e. post-edit, the missing or unsuitable information. Within the scope of this deliverable, an online platform has been developed allowing users to import automatically created dictionary drafts and post-edit them efficiently while preserving access to the underlying corpus evidence. The development was carried out within the scope of the Lexonomy [3] dictionary writing system that has been enhanced with these post-editing features.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2. Sketch Engine

access on **www.sketchengine.eu**

Sketch Engine is corpus management, corpus building and text analysis software developed by Lexical Computing (find more [1]). Originally developed for lexicography, it is now used by a variety of users such as lexicographers, researchers in corpus linguistics, translators, interpreters, language teachers, language learners and others in need of understanding how language is used. Sketch Engine currently contains corpora in 90+ languages and supports user corpus building in all of them. The largest corpora consist of texts in the total length of 40 billion words and their size grows daily. Some of the corpora are the largest available corpora in the language.

Sketch Engine is a complex suite of a variety of tools designed for searching effectively large text collections of billions of words according to complex and linguistically motivated queries. Sketch Engine is designed with a special emphasis on scalability and search speed.

OneClick Dictionary – The idea behind the OneClick Dictionary tool consists in the belief that dictionary making and dictionary editing could be much more productive, faster and cheaper if dictionary entries were pre-generated automatically with data coming from text corpora (Figure 4). Such dictionary drafts would still need to be post-edited by lexicographers but deleting, amending and rephrasing is more productive than developing dictionary entries from scratch. OneClick Dictionary triggers all the Sketch Engine tools and produces a list of the **most frequent** words (using Wordlist) or the list of the **most typical** words (using Keywords & Terms). It also adds information about the most typical **collocations** (using Word Sketch), **example sentences** (using the concordance with GDEX), **translations** (using parallel corpora), **synonyms** (using Thesaurus), **word forms**, **part of speech** or **definitions**. The user can also activate automatic word sense disambiguation. The final database of dictionary entries is automatically pushed to Lexonomy [3] for post editing.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

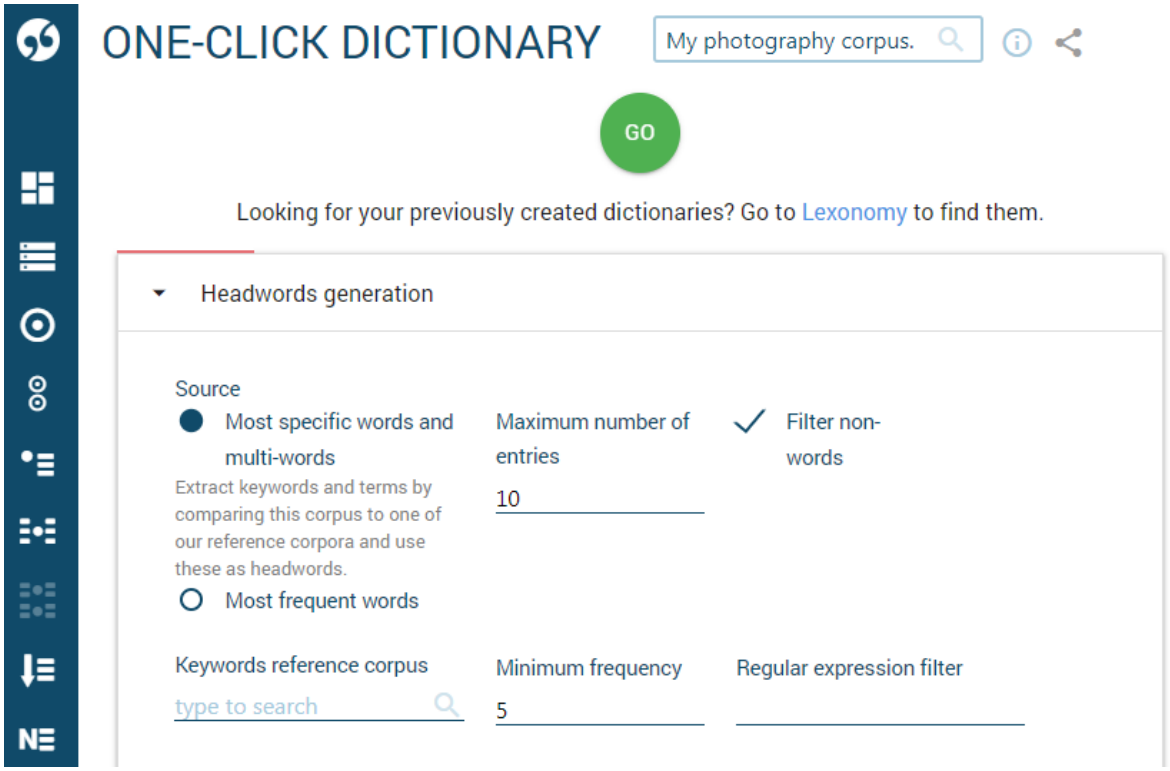


Figure 1. OneClick Dictionary – setting up the building of a new dictionary draft from a corpus.

OneClick Dictionary is not limited to professional lexicography but is also designed for spontaneous lexicography – small projects of lexicographic nature such as glossaries and domain-specific wordlists and dictionaries often prepared by teachers or other professionals without formal training in lexicography. Such projects are numerous at various academic and educational institutions and the OneClick Dictionary tool will provide the needed support and simplicity.

A more detailed description of Sketch Engine can be found in the Deliverable D4.1 Online Dictionary Post-Editing and Presentation Module

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

3. Lexonomy

access on **www.lexonomy.eu**

Lexonomy is a cloud-based open-source dictionary writing and online dictionary publishing system (see more in [3]) which is highly scalable and can adapt to large dictionary projects as well as small lexicographic works such as editing and online publishing of domain-specific glossaries, wordlists or terminology resources. Lexonomy allows editing from scratch but also accepts automatically generated dictionary drafts **pushed** to Lexonomy from Sketch Engine via a dedicated connection. During the editing process, users can also **pull** data from the corpora in Sketch Engine whenever they are needed during the entry editing process. The final dictionary can be exported or simply published online, accessible via a dedicated link in a desktop and mobile-friendly (Figure 2) user interface.

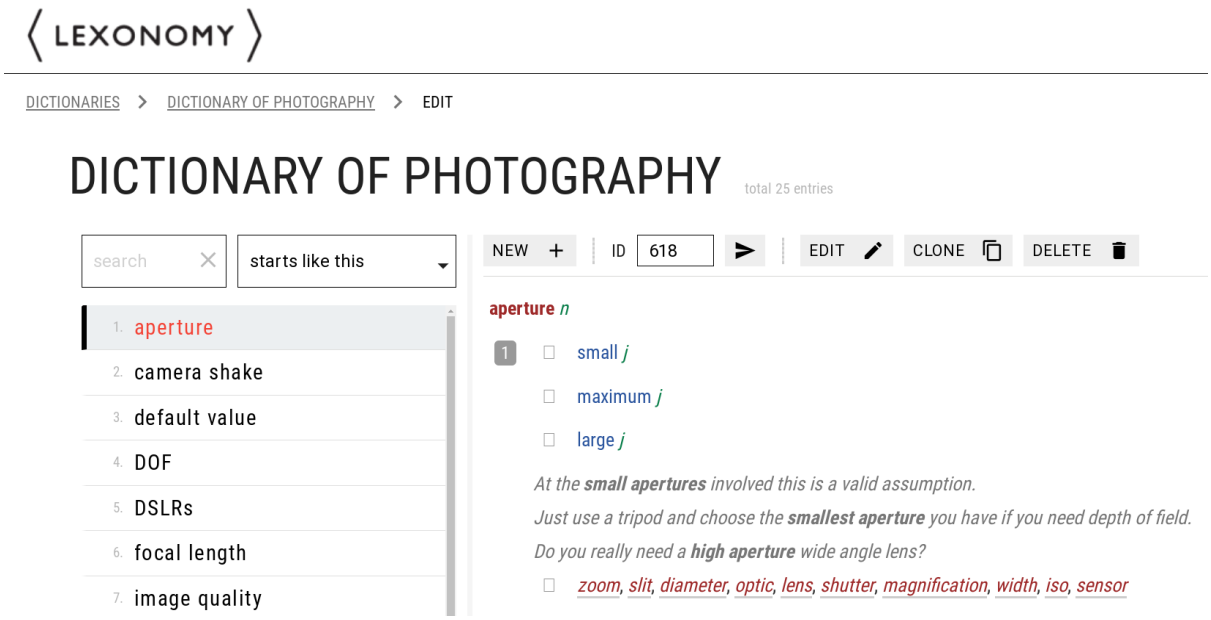


Figure 2. A dictionary entry within Lexonomy.

A more detailed description of Lexonomy can be found in the Deliverable D4.1 Online Dictionary Post-Editing and Presentation Module

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

4. Experiment description

The aforementioned tools were used in the context of a commercial lexicographic project to create three bilingual dictionaries from scratch. The source languages of those dictionaries were Lao, Urdu and Tagalog, with target languages being English and Korean. We first crawled web corpora for the respective three source languages according to a procedure described in [4].

Dictionary composition and entry structure

The goal was to create a dictionary of 50,000 headwords, out which the 15,000 most frequent one (according to the document frequency) would be manually post-edited.

The structure of each dictionary entry was as follows:

- headword list
- inflected forms
- audio pronunciation
- for each sense
    - a sense disambiguator
    - 1-10 collocations per sense
    - 1-10 synonyms/antonyms per sense
    - 1 picture per sense (where appropriate)
    - 3 example sentences per sense
    - English translation of the sense disambiguator and 1 example per sense
    - Korean translation of the sense disambiguator and 1 example per sense

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

5. Source corpora

The overall statistics for the corpora is described in Table 1.

language	corpus	number of tokens	number of unique word forms	number of unique lemmas
Tagalog	tlTenTen19	198M	3,006,551	2,225,117
Lao	loTenTen19	105M	874,599	-
Urdu	urTenTen18	273M	5,301,083	1,726,019

Table 1. Corpus statistics for the web corpora used for dictionary drafting

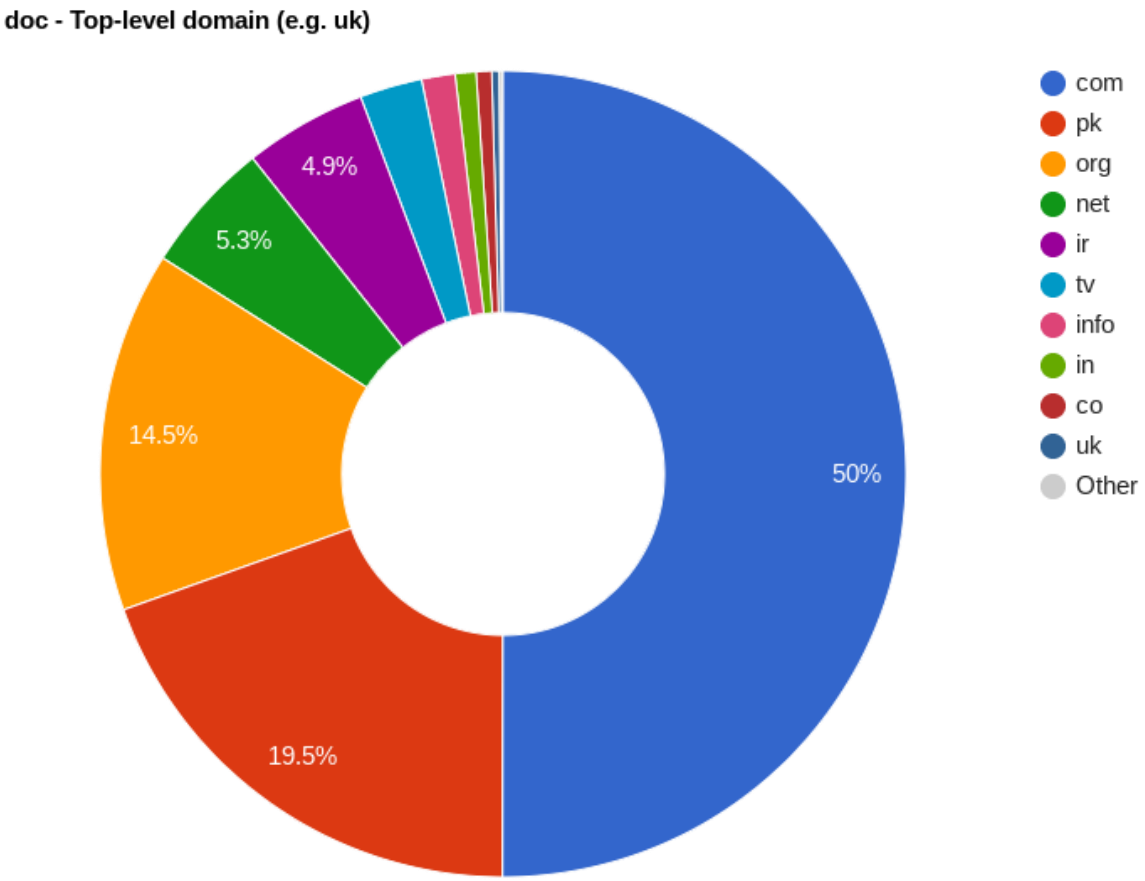


Figure 1: Top-level domain names for the urTenTen18 Urdu corpus

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Corpus sources

The corpora were crawled by means of a general web crawl using the Spiderling crawler [10], and then cleaned and deduplicated using the Justext and Onion tools [11]. The corpus composition as for top-level domain names is provided in Figures 1–3.

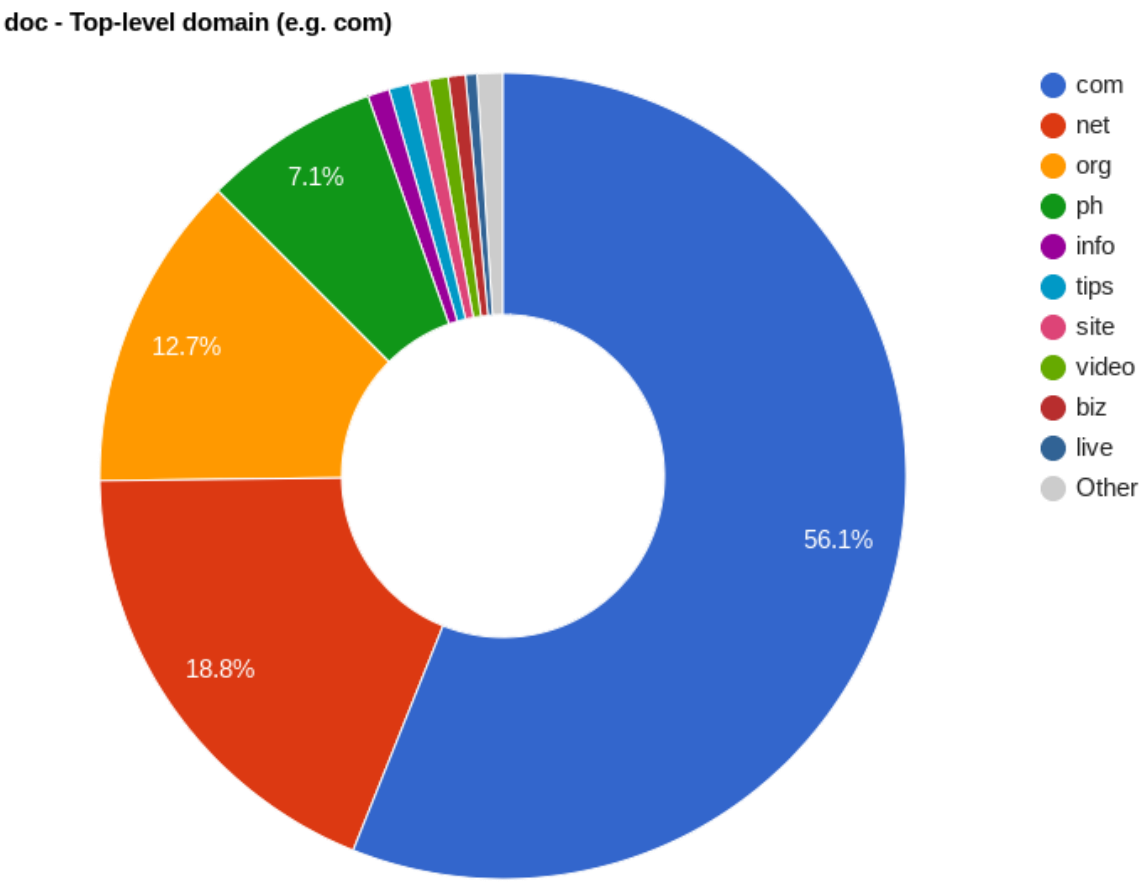


Figure 2: Top-level domain names for the tlTenTen19 Tagalog corpus

Corpus annotation

These corpora were part-of-speech tagged and (where necessary – Lao is not a flective language) lemmatized.

1. **Tagalog**

   We used a modified version of the freely available Stanford parser for tagging as trained in [5] and significantly expanded version of a free lemmatizer available in [6].

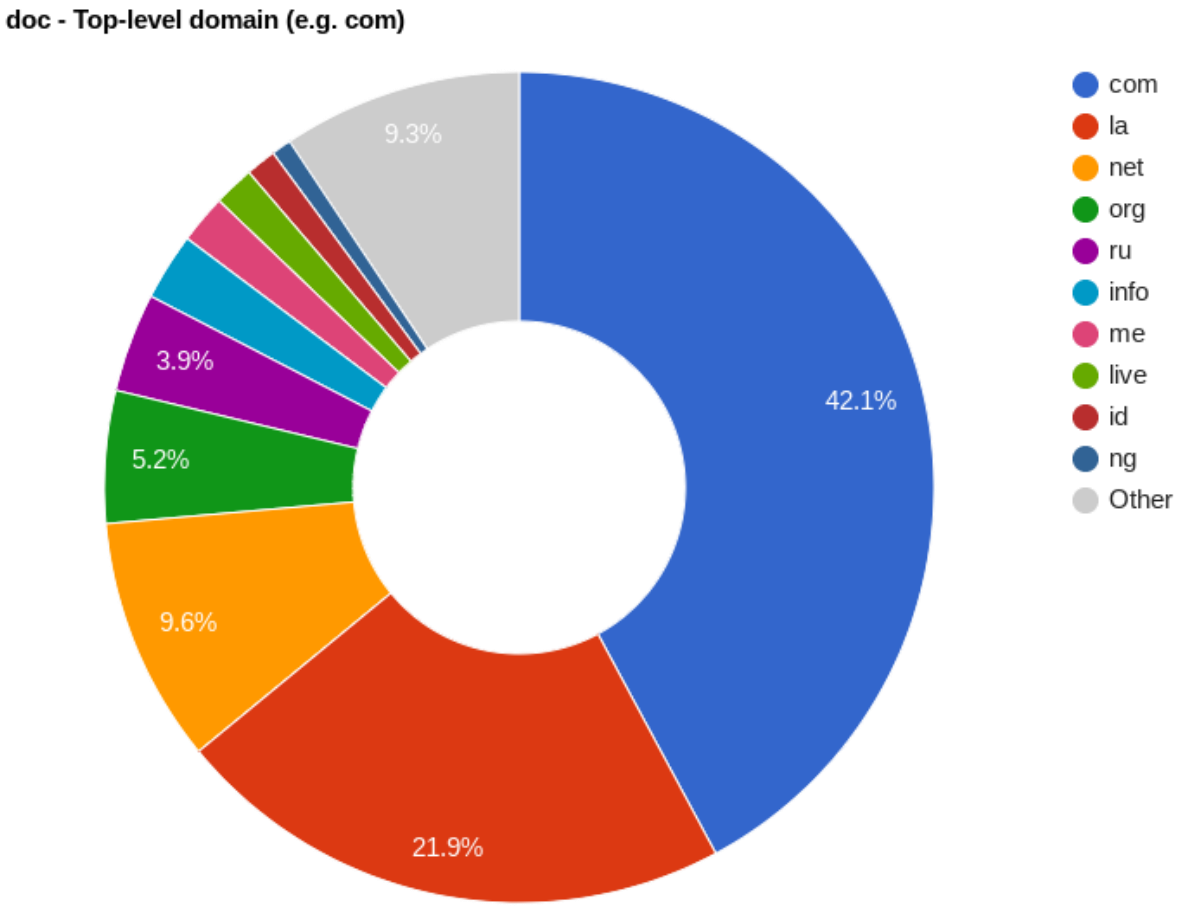
This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

2. **Urdu**

   The Urdu corpus was initially part-of-speech tagged and lemmatized using the IIIT Hyderabad parser [7] and then further on improved using RFTagger [8] trained on the Urdu Universal Dependency Treebank dataset [9] (part of the Universal Dependencies project[1]).

3. **Lao**

   The Lao corpus was part-of-speech tagged using RFTagger [8] based on a model we trained on the PANL10N Lao corpus.



[1] https://universaldependencies.org/

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Figure 3: Top-level domain names for the loTenTen19 Lao corpus

6. Post-editing workflow

The overall post-editing workflow is presented in Figure 4: having the corpus we first automatically generated the headword list which allowed us to automatically generate the list of inflected forms (based on the lemmatization of the corpus) and perform automatic word sense induction. We also recorded audio pronunciation (this step was not automated and post-edited, for obvious reasons). After the word senses were post-edited, we automatically generated example sentences, thesaurus and downloaded images from the web. Finally we performed the translation tasks.

Each of the steps was implemented as a standalone dictionary in Lexonomy (representing a batch to be post-edited) equipped with a custom editing widget. In [12] a detailed description of each post-editing step can be found.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

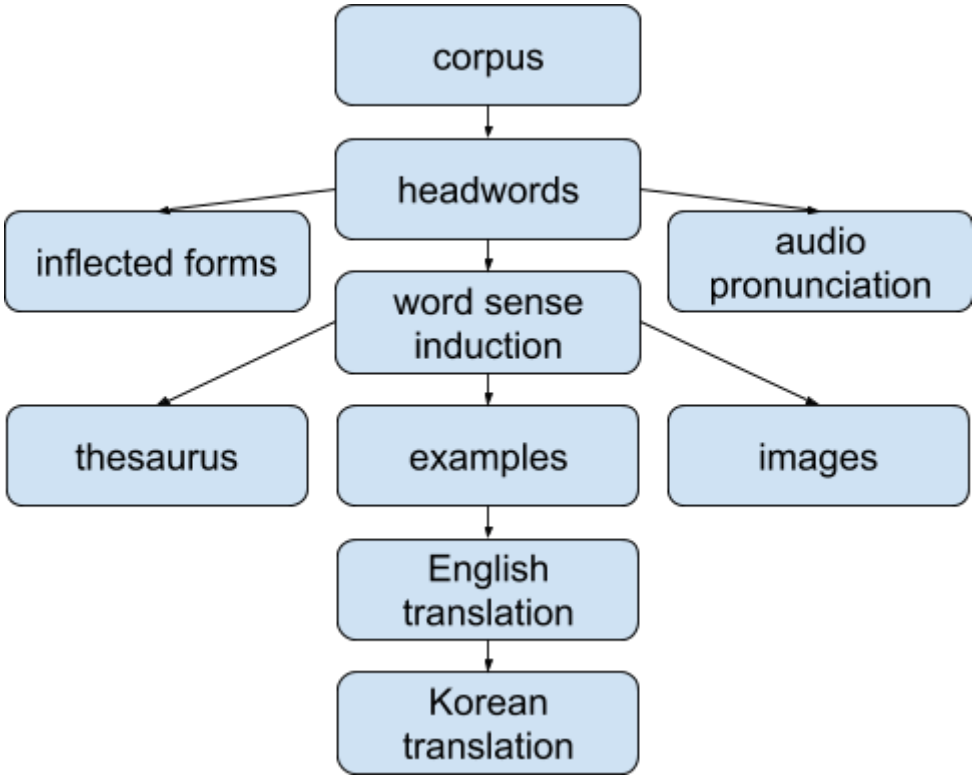


Figure 4: Post-editing workflow used in the evaluation

A batch was initially never edited only by one annotator but multiple (typically five or more) were used and inter-annotated (IAA) agreement was measured. Single-editing of batches was only carried out after all the annotators reached sufficient IAA (depending on the task), became familiar enough with the guidelines and the guidelines were adjusted following the issues observed in the initial multi-annotated batches.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

7. Conclusions

From the experiments performed it follows that the tools and methods developed as part of the D4.1 Online Dictionary Post-Editing and Presentation Module and D4.2 Dictionary Drafting Module can be successfully deployed for building large dictionaries completely from scratch. In [12] we ellaborate in more detail on the implications which can be summarized in the sense that the methodological changes and issues turned out to be much more important and substantial than the technological ones.

In other words, the technology is ready and its performance is sufficient to make the post-editing approach viable and efficient, alas the methodology not so much. The process is quite different from a traditional lexicographic workflow focusing on editing the whole entry (with subsequent reviews) and has many implications for the lexicographic judgments made, some of which are yet to be discovered.

Despite the challenges in methodology and human/data management, this approach enables lexicographers to produce dictionaries faster – thanks to the automation – and better – thanks to the fact many of the tasks can be delegated to educated native speakers, whereas senior lexicographers can focus on the most demanding lexicographic judgments and supervision.

References

[1] KILGARRIFF, Adam, Vít BAISA, Jan BUŠTA, Miloš JAKUBÍČEK, Vojtěch KOVÁŘ, Jan MICHELFEIT, Pavel RYCHLÝ and Vít SUCHOMEL. The Sketch Engine: ten years on. In *Lexicography*. Berlin: Springer Berlin Heidelberg, 2014, p. 30–34.

[2] Rundell, M. (2008). The corpus revolution revisited. English Today, 24(1), 23-27. doi:10.1017/S0266078408000060

[3] MĚCHURA, Michael Boleslav. Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*. Brno: Lexical Computing CZ s.r.o., 2017, p. 19–21.

[4] JAKUBÍČEK, Miloš, et al. The TenTen corpus family. In: *7th International corpus linguistics conference CL*. Lancaster University, 2013. p. 125-127.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

[5] Nocon, N. and Borra, A.'s "SMTPOST: Using Statistical Machine Translation Approach in Filipino Part-of-Speech Tagging" (2016) from De La Salle University, Manila, Philippines.

[6] Carl Jerwin F. Gensaya, Tagalog Words Stemmer using Python, Available from: https://github.com/crlwingen/TagalogStemmerPython

[7] Department of Information Technology Ministry of Communications & Information Technology Govt. of India, Unified Parts of Speech (POS) Standard in Indian Languages - Draft Standard –Version 1.0. Available from: http://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf

[8] Schmid, H. and Laws, F. 2008. Estimation of conditional Probabilities with Decision Trees and an Application to Fine-Grained POS tagging, COLING 2008, Manchester, Great Britain.

[9] Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. The Hindi/Urdu Treebank Project. In the Handbook of Linguistic Annotation (edited by Nancy Ide and James Pustejovsky), Springer Press

[10] J. Pomikalek and V. Suchomel 2012. Efficient Web Crawling for Large Text Corpora Proc. 7th Web-as-Corpus workshop, Lyon, France.

[11] J. Pomikalek 2011. Removing Boilerplate and Duplicate Content from Web Corpora. PhD thesis, Masaryk University, Brno, 2011.

[12] BLAHUŠ, Marek, et al. Semi-automatic building of large-scale digital dictionaries. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, 99.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.