D4.8 Evaluation and assessment of methods for automatic enriching of lexicographic resources

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

D4.8 Evaluation and assessment of methods for automatic enriching of lexicographic resources

Deliverable Number: D4.8
Dissemination Level: Public
Delivery Date: July 31, 2022
Version: 1

Authors: 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.

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 describes the testing results of the features newly implemented within the Dictionary Enhancement Module (D4.4) in the ELEXIS infrastructure. We describe the experiments performed, mostly in the production environment, and the results achieved by the respective parts of the Dictionary Enhancement 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.

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 analyse 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

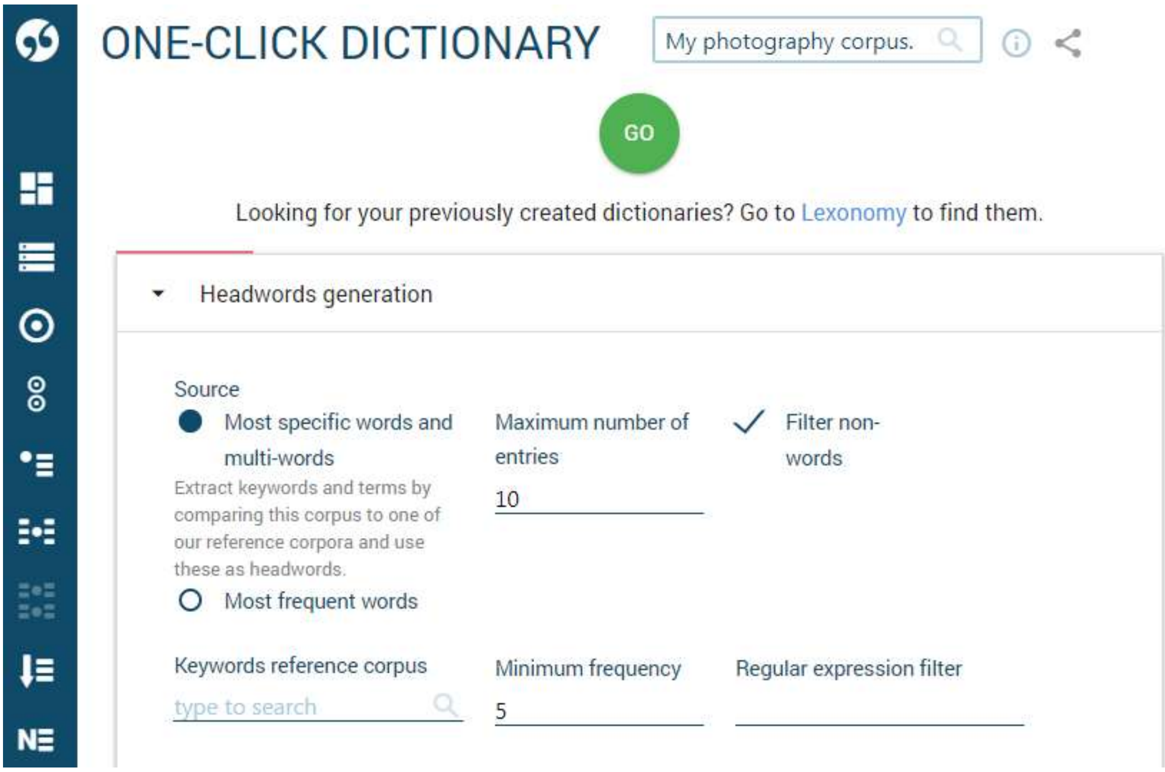
access on **www.sketchengine.eu**

Sketch Engine is corpus management, corpus building and text analysis software developed by Lexical Computing (find more in [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.



ONE-CLICK DICTIONARY

My photography corpus.

GO

Looking for your previously created dictionaries? Go to Lexonomy to find them.

Headwords generation

Source

● Most specific words and multi-words

Extract keywords and terms by comparing this corpus to one of our reference corpora and use these as headwords.

○ Most frequent words

Maximum number of entries: 10

✓ Filter non-words

Keywords reference corpus: type to search

Minimum frequency: 5

Regular expression filter:

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 word lists 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.

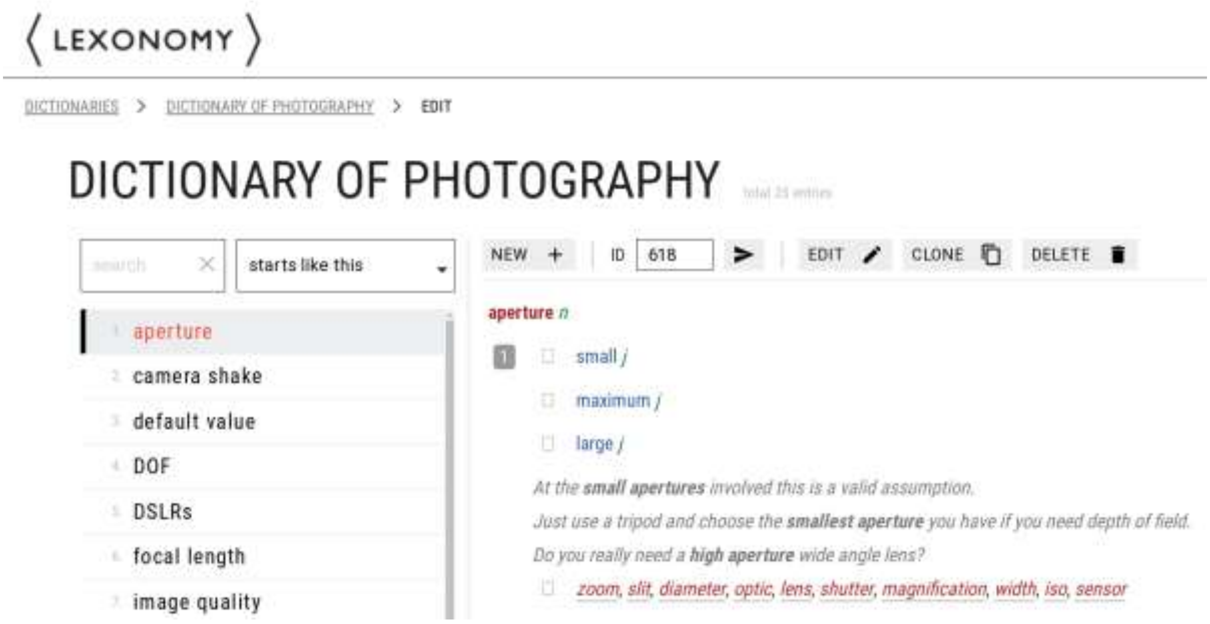
A more detailed description of Sketch Engine can be found in the Deliverable D4.1 Online Dictionary Post-Editing and Presentation Module.

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 10) user interface.



DICTIONARY OF PHOTOGRAPHY total 25 entries

1 aperture
2 camera shake
3 default value
4 DOF
5 DSLRs
6 focal length
7 image quality

aperture *n*

1 □ small /
□ maximum /
□ large /

*At the **small apertures** involved this is a valid assumption.*
*Just use a tripod and choose the **smallest aperture** you have if you need depth of field.*
*Do you really need a **high aperture** wide angle lens?*

□ *zoom, slit, diameter, optic, lens, shutter, magnification, width, iso, sensor*

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. NAISC

‘NAISC’ means ‘links’ in Irish and is pronounced ‘nashk’.

NAISC 1.0 (https://github.com/insight-centre/naisc) is a tool for linking datasets and was created by the SFI Insight Centre for Data Analytics and the ELEXIS project. NAISC serves as a system for aligning RDF datasets: It takes 2 RDF documents as input (referred to as *left* and *right*) and outputs an alignment (set of RDF triples) between these two documents. NAISC typically relies on a configuration, which is a JSON document.

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. Enrichment features

In the following we describe individual enrichment features that operate using the Sketch Engine, Lexonomy and NAISC.

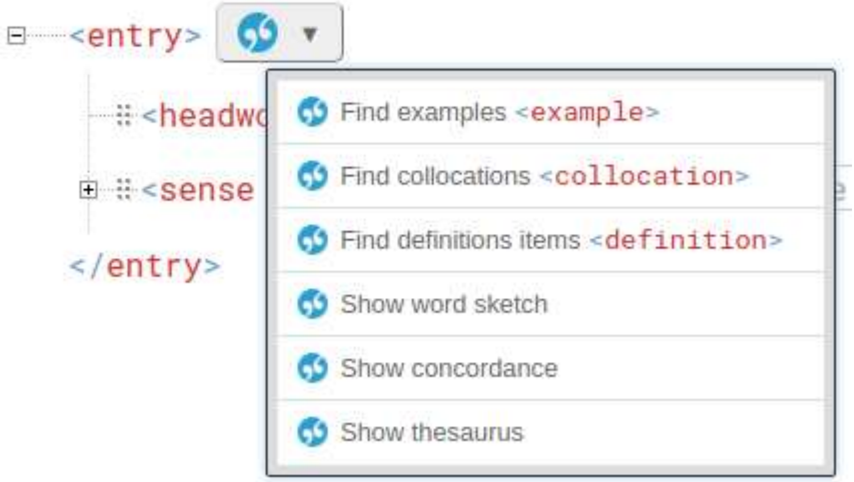


Figure 3. Menu in Lexonomy illustrating some of the Dictionary Enhancement Module features

5.1 Corpus examples

Within the work on the Dictionary Enhancement Module, we implemented automatic retrieval of the dictionary examples from a corpus. This functionality was tested in a large-scale project where 3 examples were needed for each of the 13,000 headwords in a dictionary of Tagalog (Filipino).

For each of the words, 10 corpus examples were generated using the GDEX technology (selecting best examples out of available corpus examples) built into the Dictionary Enhancement Module, out of which three were selected by the editors, then post-edited and included in the dictionary. We have measured the number of examples accepted without any changes and the number of examples accepted with only minor changes.

In total, 130,000 examples were automatically generated from the corpus, out of which 39,000 were selected for the dictionary. Of these 39,000 examples, 17,935 (46%) were accepted as they were, without any change. Another 3,779 examples (10%) required only minor changes. The rest (17,286 examples, i.e. 44%) required more 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.

Selecting examples was roughly 4 times faster than editing or creating them, so nearly 50% of the resources for the project were saved by using the new functionality. In addition, the examples generated from the corpus are more natural and closer to authentic language.

Although the numbers above indicate significant improvements, we originally expected them to be even higher. Therefore, we investigated the reasons why so many examples required so much post-editing, and came to the conclusion that the corpus size is an important factor in the quality of extracted candidate examples. In other words, the bigger the corpus, the better the selection of candidates. The Tagalog corpus has less than 200 million words, so the success was somewhat limited.

To further support this hypothesis, we performed another experiment with the same setting (selecting 3 good examples out of 10 candidates) and with a 1-billion-word clean web corpus of English. The scale of the experiment was much smaller, only 100 random words were assessed, but the results seem to confirm the hypothesis: Out of the 300 selected example candidates, only 14 examples needed significant post-editing and further 13 required minor changes. So, with a large clean corpus, this feature will lead to large savings in the dictionary creation process.

5.2 Corpus collocations

Similarly to the example extraction, the new methods enable automatic retrieval of collocation candidates for each word. This feature was tested on a large scale (collocations were added semi-automatically for about 13,000 headwords in the dictionary) with three very different languages (Tagalog, Urdu and Lao). We have measured the total number of collocations generated, and the number of collocations marked as correct by the dictionary editors. The results are as follows:

- Tagalog: 216,806 collocations generated, 103,877 marked correct (48%)
- Urdu: 235,716 collocations generated, 125,038 marked correct (53%)
- Lao: 285,737 collocations generated, 194,669 marked correct (68%)

Again, all the three corpora were rather small by today's standards (197, 240 and 105 million words, respectively) and it can be expected the numbers will rise with larger corpora. However, even the approximately 50% success rate indicates that the new feature is extremely useful – as most dictionaries don’t contain rich information about collocations, and the new feature significantly streamlines the process of adding collocations into the dictionary.

5.3 Thesaurus

The Dictionary Enhancement Module enables lexicographers to extract similar words according to the distributional thesaurus generated from a corpus. This new feature was tested again with the projects involving the Tagalog, Urdu and Lao dictionaries (each with about 13,000 entries). In total, 244, 369 and 254 thousand candidates were generated, and the following numbers of similar words, synonyms or antonyms were added:

- Tagalog: 8,529 synonyms, 1,460 antonyms and 4,580 similar words
- Urdu: 16,484 synonyms, 4,253 antonyms and 33,580 similar words
- Lao: 14,408 synonyms, 2,439 antonyms and 21,891 similar words

In total, it means 1.2–4.2 semantically related words per headword. Such an amount of related words would be unrealistic to achieve in a reasonable time without the new automatic corpus methods.

5.4 Corpus definitions

Adding corpus definitions into the dictionary has not yet been used in any major project, it was just tested experimentally. From these initial tests it seems that corpora of general language do not contain many quality definitions suitable for dictionaries; however, there may be cases where there are domain specific data (e.g. for domain specific dictionaries), that have to contain the definitions of the specialised vocabulary. Unfortunately, no such data and no such project was undertaken yet, so this part of the Dictionary Enhancement Module stays so far in a rather experimental phase, but the technology is there, ready for its first production use.

5.5 Audio pronunciation

Accuracy, and thus the usefulness of the automatic audio synthesis heavily depends on the language – for English and many other major languages, the technology is practically perfect, whereas for some low-resourced languages speech synthesis is not even available. An evaluation on particular languages within the VoiceRSS API would therefore be not really informative.

Also, compared to the previously mentioned tasks (examples, collocations, thesaurus), recording audio directly by native speakers is relatively fast and cheap, provided there is reasonable sound recording equipment available. Automatic audio synthesis is therefore useful in cases where there is no native speaker available, or in the case of financial constraints and when the best quality is not an absolute value for the project.

We have used the new feature with success in several projects where the 100% quality was not a top priority (e.g. a part of the Tagalog, Urdu and Lao projects mentioned above where audio was deliberately synthesised automatically). No thorough evaluation was performed but the feedback was always positive, indicating typical accuracy around 95% or higher.

5.6 Images

To test the usability of the functionality of adding images, we added an image to each of the 22,026 senses of the 13,000 entries in a Tagalog-English dictionary, and let the dictionary editors select an appropriate image out of the menu of 10 automatically generated images. If there was no appropriate image, editors were instructed not to select anything.

Out of the 22,026 senses, the editors found an appropriate image in 18,057 cases (82%). Apart from the high accuracy of the image candidates (taking into account that for some words no appropriate image exists due to the character of the words), it is clear that such an amount of images could not be added to the dictionary in a reasonable timeframe without the specialised tool within the Dictionary Enhancement Module.

5.7 Linking entries

Automatic linking with the NAISC tool, built into Lexonomy within the scope of the Dictionary Enhancement Module, supports only monolingual links. Therefore, linking was only processed for languages where multiple dictionaries are available in the Elexis data. The following table summarises the results of automatic linking for each dictionary pair.

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.

Language	Dictionary 1	Dictionary 1, number of entries	Dictionary 2	Dictionary 2, number of entries	Number of links
Danish	The Danish Dictionary	5100	Dictionary of the Danish Language	4500	1409
Danish	The Danish Dictionary	5100	Moth's Dictionary	93832	205
German	Middle High German Dictionary	36048	The Schweizerisches Idiotikon	160254	1257
German	Schranka 1905 - Wiener Dialekt-Lexikon	1334	The Schweizerisches Idiotikon	160254	89
Bulgarian	BG Dictionary Synonyms	29998	BG Explanatory Dictionary	59622	16428
Slovenian	JSV	8461	Pletersnik Dictionary	103185	4030
Slovenian	JSV	8461	CJVT Thesaurus	105473	3207

As seen from the results, the NAISC success rate is dependent on the language of the dictionary. For example, the dictionaries “Schranka 1905” and “The Schweizerisches Idiotikon” both describe regional dialects of German, so the overlap and number of possible links between them are low. On the other hand, Slovenian or Bulgarian dictionaries describe the same language, so the number of automatically produced links is much higher, with a recall rate of 54%.

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.

Conclusions

The individual components of the Dictionary Enhancement Module were designed according to practical needs of dictionary editors, to streamline their work and make it easier and more efficient, from the very beginning. In this document, we described real-world experiments, mostly in real-time production use cases, which illustrate that the module really performs what it was designed for; with the newly developed tools, creating rich dictionary entries is faster and easier, and the entries are also likely to contain higher-quality data compared with older dictionaries, thanks to their direct connection to the corpus data.

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, Michael (2008). The corpus revolution revisited. English Today, 24(1), 23-27. doi:10.1017/S0266078408000060

[3] MĚCHURA, Michal 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.

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.