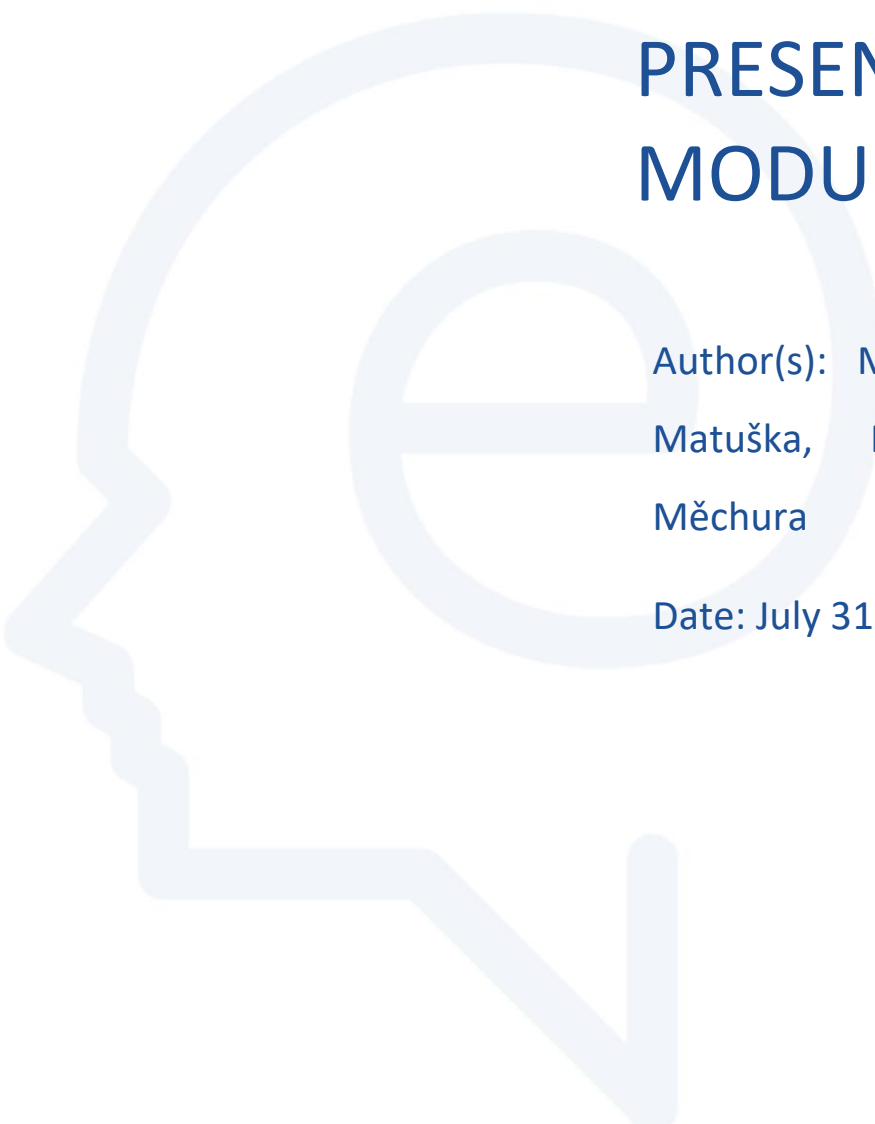


## D4.1

# ONLINE DICTIONARY POST-EDITING AND PRESENTATION MODULE



Author(s): Miloš Jakubiček, Ondřej  
Matuška, Michal Cukr, Michal  
Měchura

Date: July 31st, 2019

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D4.1 ONLINE DICTIONARY POST-EDITING AND  
PRESENTATION MODULE

Deliverable Number: D4.1

Dissemination Level: Public

Delivery Date: July 31st 2019

Version: 3

Author(s): Miloš Jakubiček, Ondřej  
Matuška, Michal Cukr,  
Michal Měchura



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 15th	Initial draft	Miloš Jakubiček
2, July 20th	Post-editing features	Ondřej Matuška
3, July 27th	Assessment by	Simon Krek





## Table of Contents

1	Introduction .....	2
2	Background: dictionary post-editing.....	3
3	Sketch Engine.....	4
4	Lexonomy.....	6
5	Dictionary post-editing .....	9
6	Dictionary presentation .....	13
7	References .....	16

## List of Figures

Figure 1: SketchEngine access on sketchengine.eu .....	4
Figure 2: OneClick Dictionary - setting up the building of a new dictionary draft from a corpus. ....	5
Figure 3: Lexonomy access on www.lexonomy.eu. ....	6
Figure 4: A dictionary entry within Lexonomy.....	6
Figure 5: Editing particular attributes of a dictionary entry within Lexonomy.....	8
Figure 6: Interlinks between dictionary entries in Lexonomy and corresponding examples from Sketch Engine.....	9
Figure 7: Lexonomy: entry lay-by.....	11
Figure 8: A list of access privileges to a dictionary in Lexonomy. ....	11
Figure 9: Mobile resolution of Lexonomy.....	14
Figure 10: Lexonomy on desktop monitors. ....	15



## 1 Introduction

This report presents an overview of the software deliverable 4.1 Online Dictionary Post-Editing and Presentation Module. We briefly outline the rationale behind the tools developed, the methodology that was involved and, finally, present an overview of the functions of the software.





## 2 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.



### 3 Sketch Engine



Figure 1: SketchEngine access on 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 2). 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.

D4.1 Online Dictionary Post-Editing and Presentation Module

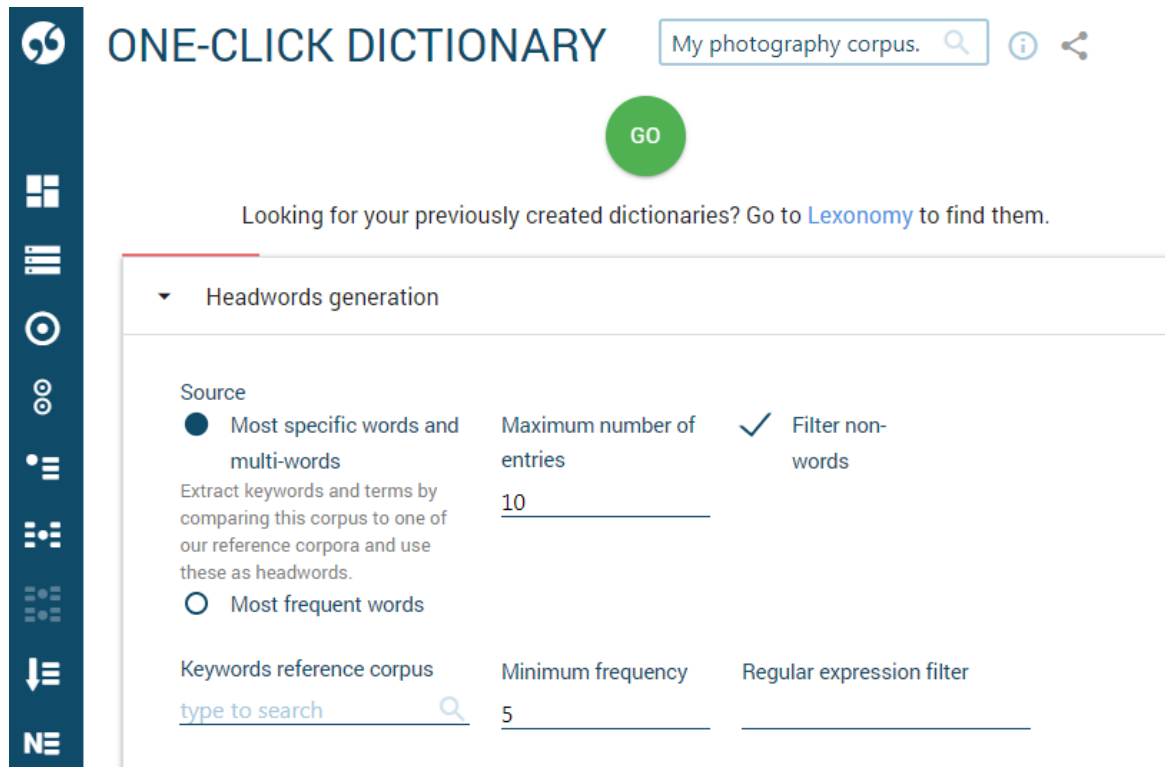


Figure 2: OneClick Dictionary - setting up the building of a new dictionary draft from a corpus.

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.



## 4 Lexonomy



Figure 3: Lexonomy access on [www.lexonomy.eu](http://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 9) user interface.

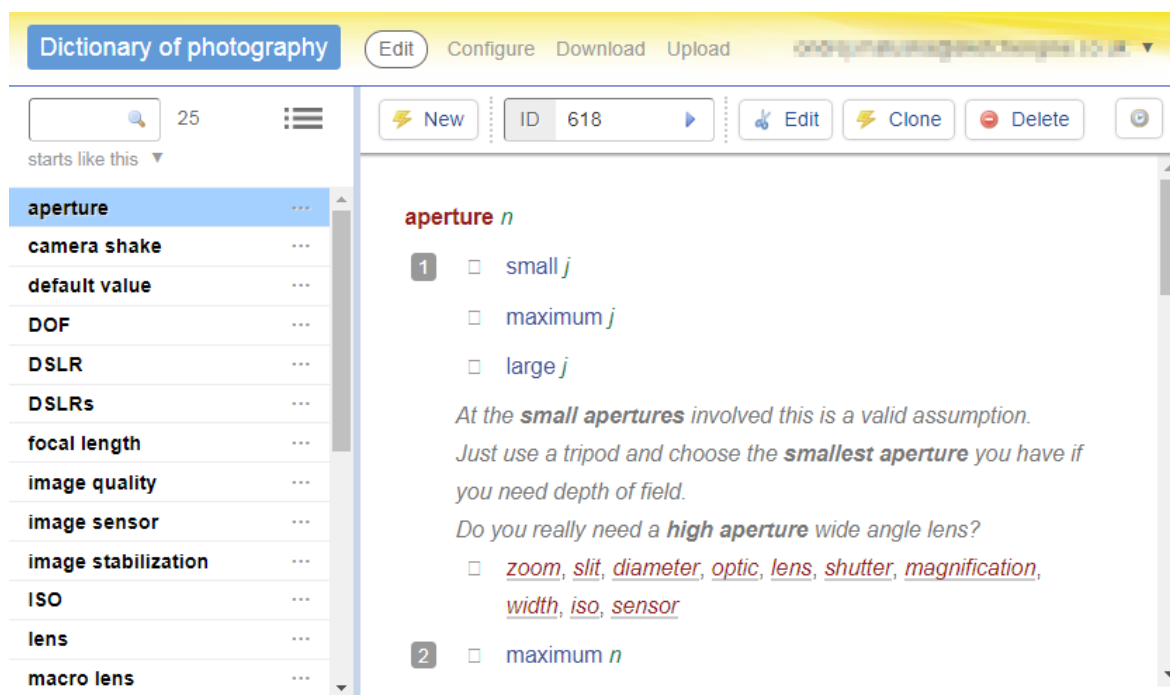


Figure 4: A dictionary entry within Lexonomy.

### **Dictionary templates**

Lexonomy supports dictionary templates which define what elements dictionary entries should or must contain. Each piece of a dictionary entry information such as pronunciation, definition, example, synonym, collocation, translation etc. can be defined as optional or compulsory, the number of such elements within the same dictionary entry can also be defined. The content of some elements can be limited to only a finite list of values such as the list of part of speech abbreviations. Any such restrictions can be defined by the user. This ensures consistency across all dictionary entries. Each dictionary template can contain an unlimited number of dictionary entry templates to accommodate different dictionary entry types. For example, dictionary entries for frequently used words with a large number of senses will have a different structure and will contain different amount and type of information than entries for rarely used words with only one sense.

### **Editing the dictionary**

The dictionary editing interface was specifically designed for users with little or no knowledge of the XML data format. [3] The interface automatically looks after the correct XML data structure (see Figure 5) and completely eliminates the error-prone procedure of typing the XML code manually. The XML elements are never typed but, instead, they are selected from a predefined list of elements. The list can be modified by the user.



#### D4.1 Online Dictionary Post-Editing and Presentation Module

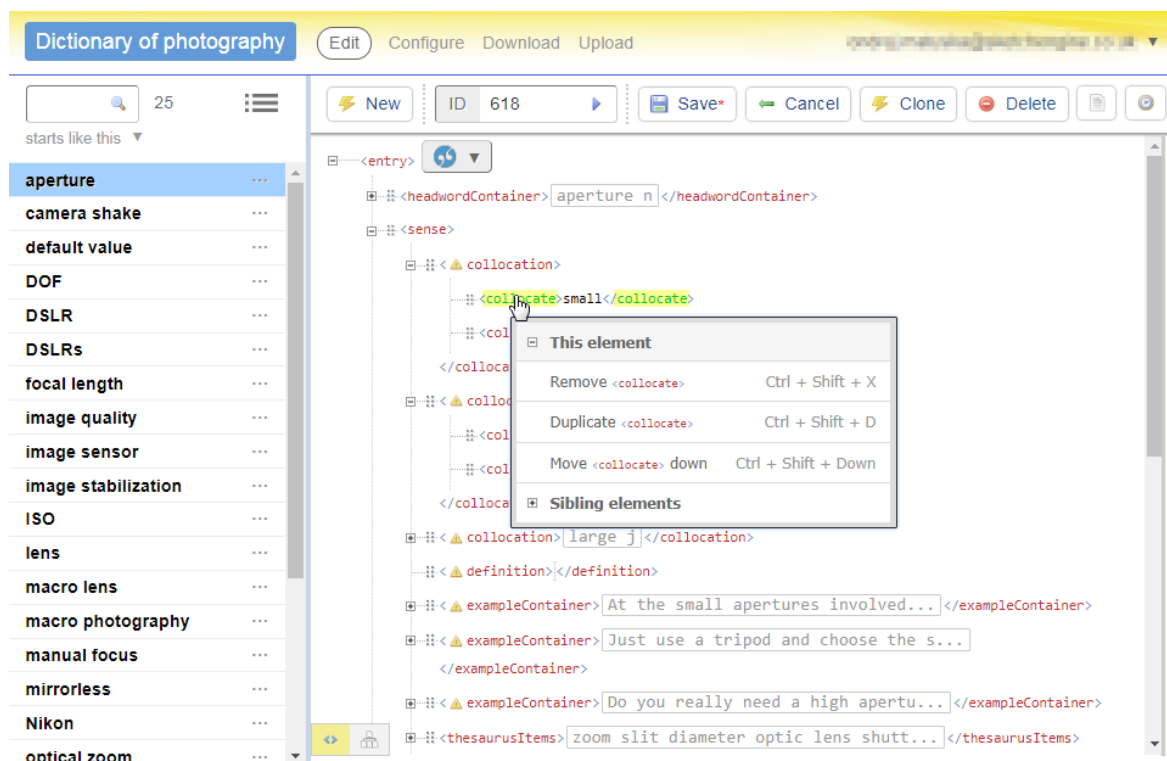


Figure 5: Editing particular attributes of a dictionary entry within Lexonomy.

Apart from operating the interface with the mouse, all editing features are also accessible via the keyboard for greater productivity.



## 5 Dictionary post-editing

### Pull model

The post-editing process may require the access to the underlying linguistic evidence in the corpus in Sketch Engine. The lexicographer may also need to pull or retrieve more data from the corpus into the dictionary entry. Lexonomy addresses these needs with a pull model, i.e. a procedure of accessing corpus data from the Lexonomy interface instead of leaving the Lexonomy interface and switching to Sketch Engine.

A dictionary draft in Lexonomy can be linked to a corpus in Sketch Engine which should be used to retrieve additional data or to check the usage in the authentic texts in the corpus. Each dictionary project can be linked to a different corpus in Sketch Engine to acknowledge the fact that a domain specific glossary might need to draw data from a different data source (corpus) than a general language dictionary.

Figure 6 shows the Sketch Engine tools which are accessible directly from the Lexonomy interface. When in editing mode, there is a Sketch Engine button (1) at the top of the entry with access to the following options: (2) tools to retrieve additional data from a corpus in Sketch Engine without leaving the Lexonomy interface and (3) links leading to the relevant section of the Sketch Engine interface where the full suite of search and analytical options is available for advanced and specific queries.

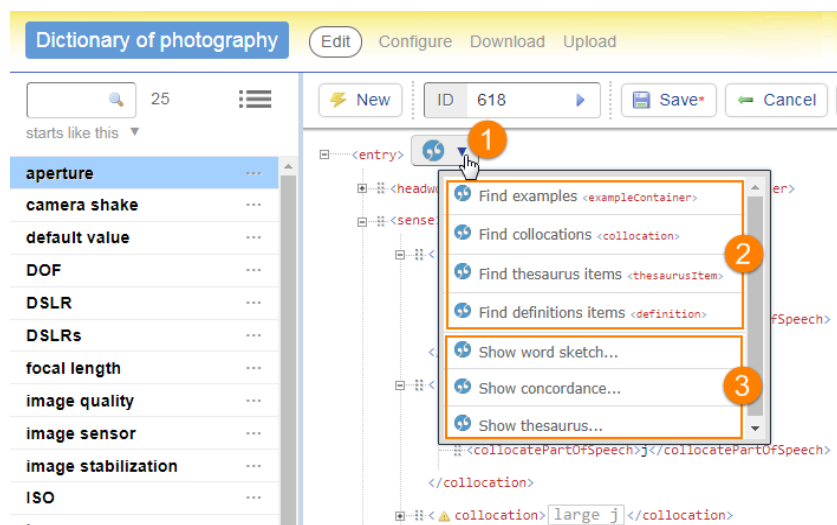


Figure 6: Interlinks between dictionary entries in Lexonomy and corresponding examples from Sketch Engine.

---

#### D4.1 Online Dictionary Post-Editing and Presentation Module

**Find examples** – This option will generate a concordance of the headword and will use the GDEX technology to evaluate the sentences according to their suitability to serve as dictionary examples. The most suitable sentences will be offered to the user who can select the ones that should be imported into the dictionary entry. Sentences are evaluated on their length, the difficulty of words (word frequency), the inclusion of brand names, proper nouns, references outside of the sentence and other expressions and linguistic features which might prevent understanding the meaning of the sentence when used outside the original context. The sentences are also evaluated with respect to ideologies or culturally sensitive topics.

Lexonomy contains two modes of searching for examples: automatic - this option will find all sentences containing the headword without the user having to type the search word, and CQL - this option makes use of the Corpus Query Language allowing the user to specified detailed criteria including parts of speech or morphological and grammatical criteria. The CQL also allows searching for lexical or grammatical structures by using linguistic parameters without specifying concrete words.

**Find collocations** – This option will generate the most typical collocations of the headword using the words sketch in Sketch Engine and will present a list of collocations for inclusion into the dictionary entry.

**Find thesaurus items** – This option will generate a list of thesaurus entries which might be synonyms or words belonging to the same semantic field. The ones selected by the user will be imported to the dictionary entry.

**Find definitions** – This option will generate a concordance of the headword and will evaluate the sentences according to their suitability to serve as definitions of the headword. The user should then select the ones to be imported into the entry.

All data pulled from Sketch Engine into Lexonomy during the post-editing process are first placed into a **lay-by** on the right of the screen (Figure 7). The lay-by can be used for storing any elements of the dictionary entry for later use or for revision by the editor before including them into the dictionary entry. The elements can be drag-and-dropped from the lay-by to the entry or from the entry to the lay-by.





#### D4.1 Online Dictionary Post-Editing and Presentation Module

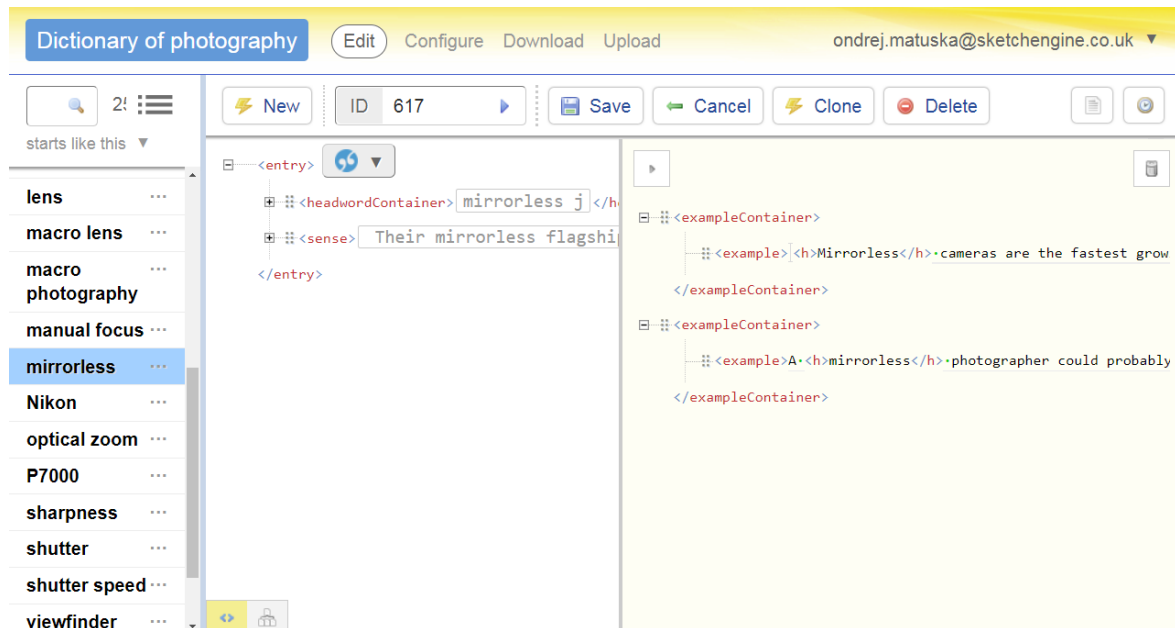


Figure 7: Lexonomy: entry lay-by.

### Collaborative editing

Lexonomy supports collaborative work which is vital to lexicographic projects. Additional users can be granted access to the unpublished dictionary.

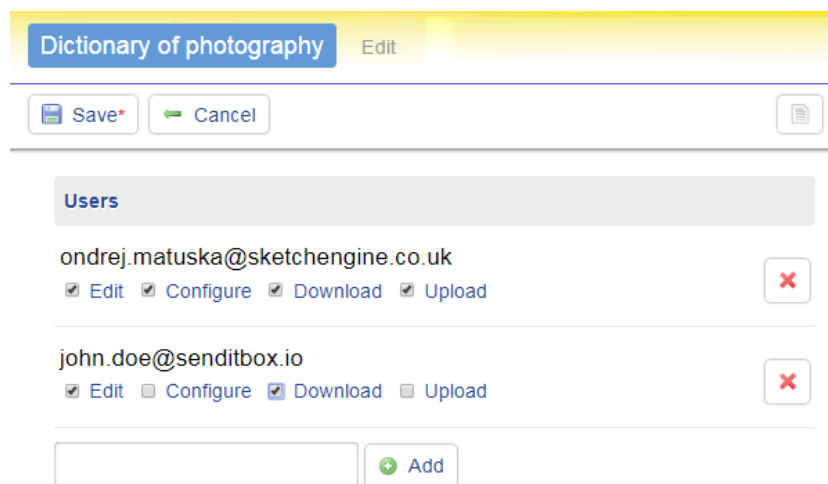


Figure 8: A list of access privileges to a dictionary in Lexonomy.



#### D4.1 Online Dictionary Post-Editing and Presentation Module

Different levels of access can be assigned as seen in Figure 8:

**Edit** - the user can make changes to the content of the dictionary entries. This setting is useful for dictionary editors.

**Configure** - the user can edit the configuration of the dictionary. This setting is for project managers or team leaders.

**Download** - this setting allows the user to download the content of the dictionary in a format that allows importing the content into various types of other software.

**Upload** - this setting allows the user to upload additional data into the dictionary.

These settings do not affect who has access to the dictionary after the dictionary is published.



## 6 Dictionary presentation

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

Lexonomy is also designed as a dictionary publishing system allowing to immediately publish the dictionary content online and also configure the appearance of each element within the dictionary entries.

### Dictionary visualisation

Each element within the dictionary entry has its default styling (colour, font size, the use of italics or boldface). The user can, however, override this styling to adapt it to the concrete lexicographic project. Custom dictionary entry elements can be defined to accommodate complex formatting of various types of entries long entries with multiple senses or sections. Lexonomy also supports conditional formatting and scripting which can be used to automatically adapt the visualisation of individual entries depending on the data they contain. For example, if the number of senses of an entry exceeds a certain limit, a word sense disambiguation menu will be shown at the beginning of the entry while it will stay hidden with short entries.

### Search configuration

The dictionary interface includes search functionality whose behaviour depends on the settings in the dictionary configuration. The search can be configured to only search in headwords, to search in all elements or to define which elements should be left out of the search. Therefore it is possible to configure the dictionary to only allow searching headwords and definitions but not example sentences.

### Publishing the dictionary

The dictionary can be published at any moment by changing the status from *private* to *public*. The dictionary will become available online at a dedicated automatically generated or user-defined URL. Publishing the dictionary online will present the data via a responsive web interface which adapts to the screens of mobile devices (Figure 9) as well as desktop monitors (Figure 10).



#### D4.1 Online Dictionary Post-Editing and Presentation Module

Lexonomy facilitates the distribution of the final product making it immediately accessible to the widest possible audience.



Figure 9: Mobile resolution of Lexonomy.

The data can also be downloaded in a standardized XML format suitable for processing into a print dictionary or for inclusion into another application or software

D4.1 Online Dictionary Post-Editing and Presentation Module

The screenshot shows the Lexonomy online dictionary interface. At the top, there is a navigation bar with the text '< LEXONOMY >' and a blue button labeled 'Dictionary of photography' with an 'Edit' button next to it. Below this is a search input field with a magnifying glass icon. The main content area displays the definition for 'macro lens' in red text. The definition includes a description: 'A macro lens has a reproduction ratio of 1:1 on the film or sensor plane. With small sensor format digital cameras an actual reproduction ratio of 1:1 is rarely achieved or needed to take macro photographs.' It also includes two italicized sentences: 'Most modern macro lenses use an autofocus system.' and 'Prime lenses will be significantly sharper than zoom lenses and macro lenses can be incredibly sharp.' Below the definition is a question: 'What macro lens did you use on these shots?'. On the right side, there is a vertical list of related terms: 'DSLRs', 'focal length', 'image quality', 'image sensor', 'image stabilization', 'ISO', 'lens', 'macro lens' (highlighted in blue), 'macro photography', and 'manual focus'.

Figure 10: Lexonomy on desktop monitors.



## 7 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.

