


D4.2

Dictionary Drafting Module



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

Date: 31. 1. 2020

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D4.2 Dictionary Drafting Module

Deliverable Number: D4.2

Dissemination Level: Public

Delivery Date: 31. 1. 2020

Version: 1.0

Author(s): Miloš Jakubíček, 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

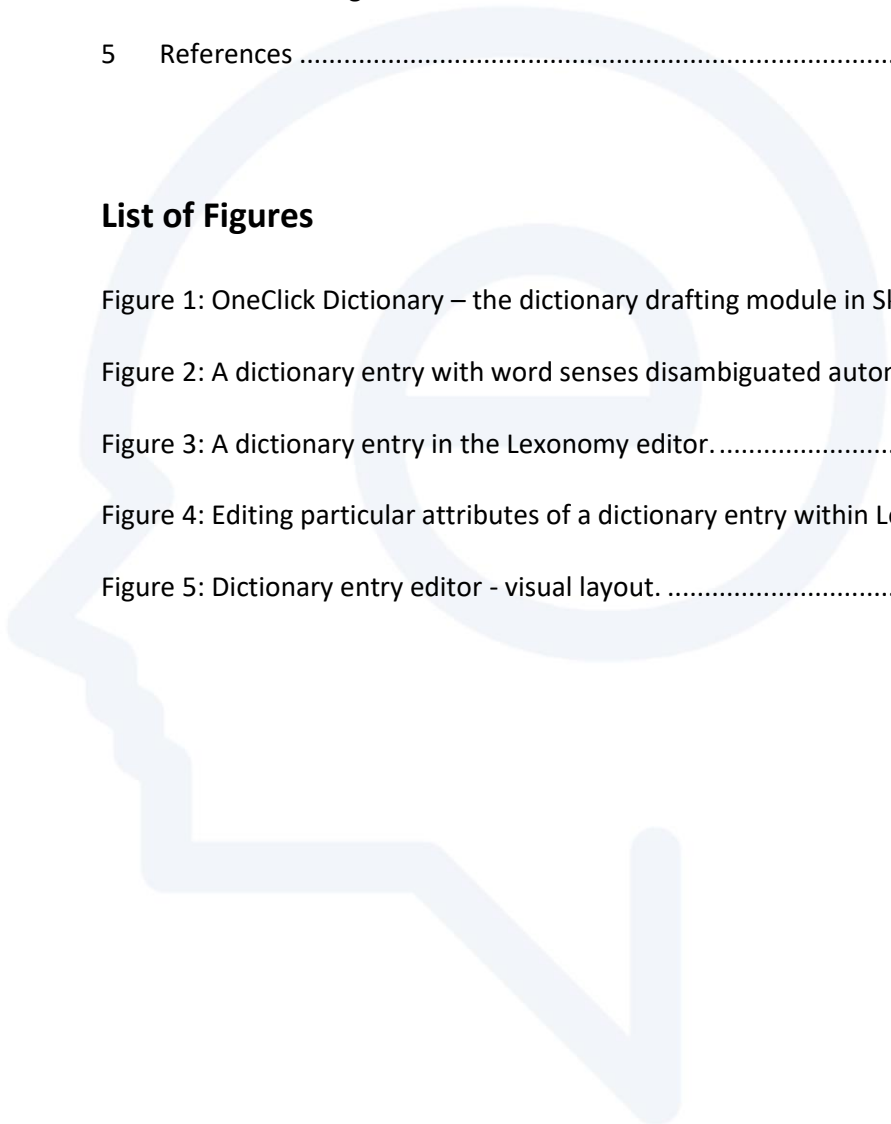
Version Date	Changes/Approval	Author(s)/Approved by
30. 12. 2019	Initial draft	Miloš Jakubíček
20. 1. 2020	Post-editing features	Ondřej Matuška
31. 1. 2020	Assessment	Simon Krek

Table of Contents

1	Introduction	1
2	Background: automatic dictionary drafting.....	2
3	Sketch Engine.....	3
4	Lexonomy.....	9
4.1	Push model	10
4.2	Pull model	10
4.3	Dictionary templates.....	10
4.4	Editing the dictionary.....	11
4.5	Dual editing interface.....	12
5	References	13

List of Figures

Figure 1: OneClick Dictionary – the dictionary drafting module in Sketch Engine.	5
Figure 2: A dictionary entry with word senses disambiguated automatically by Sketch Engine.	8
Figure 3: A dictionary entry in the Lexonomy editor.....	9
Figure 4: Editing particular attributes of a dictionary entry within Lexonomy.....	11
Figure 5: Dictionary entry editor - visual layout.	12



D6.1 Recommendations on legal and IPR issues for lexicography

1 Introduction

This report presents an overview of the software deliverable 4.2 Dictionary Drafting 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: automatic dictionary drafting

Ever since modern computers started to be used in lexicography, attempts have been made on easing the job of lexicographers and help them produce better dictionaries. Corpora have played a central role in this endeavour [2]. Initially they were used to generate headword candidates, but as bigger text corpora became available, many more new opportunities arose.

Corpora started being used extensively for the purposes of finding good dictionary examples, generating collocation candidates, calculating word similarities and finally also giving lexicographers clues about the word sense division.

The purpose of automatic dictionary drafting is to combine all these efforts and to generate a complete dictionary draft automatically. Such a dictionary draft can be used “as is” or, preferably, post-edited by a lexicographer. The post-editing phase was covered by the deliverable D4.1 Online Dictionary Post-Editing and Presentation Module and therefore it is not discussed in detail in this document.

The dictionary drafting module that was developed is effectively a wrapper around several functions of the Sketch Engine corpus management system (see Section 2) that combines them and creates a dictionary draft in the Lexonomy dictionary editor (see Section 3). The module source code is available on <https://github.com/elexis-eu/ocd>.



3 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 – OneClick Dictionary is the name of the dictionary drafting module in Sketch Engine. 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. 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.

The idea of increasing effectivity by post-editing machine-generated content rather than creating the content from scratch by a human has already been tested with translators. Many current CAT (Computer Assisted Translation) tools contain a machine translation module so that texts can be pre-translated automatically and translators then check and correct the machine generated content. This led to higher productivity and reduced costs.

The ongoing disruption in modern lexicography has been discussed by many. It is mainly two fold: the uptake of automated technologies (see e.g. Rundell and Kilgarriff, 2011) and impact of technological innovation on the publishing business (e.g. Rundell, 2013). In the former



D6.1 Recommendations on legal and IPR issues for lexicography

paper authors outline the lexicographic process when creating a dictionary as consisting of, first, corpus creation, and then headword list development, analysis of the corpus as for word senses and other lexical units and their features (collocations, colligations or preferred text types), followed by providing of definitions (or translations), corpus-based examples and final editing of the dictionary. In this paper we present a fully automated solution that implements all of the above mentioned tasks within the Sketch Engine corpus management system (Kilgarriff, 2014) and Lexonomy dictionary writing system (Mechura, 2017). The ultimate goal of this methodology is to shift all lexicographers work and intellectual input into the post-editing phase instead of manual analysis of input data before creating a dictionary draft. The connection between Sketch Engine and Lexonomy is bidirectional through a push and pull model in the sense that it allows for easy access to the corpus evidence from Lexonomy.

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 provides the needed support and simplicity.

OneClick Dictionary triggers all Sketch Engine tools relevant to dictionary building and generates a dictionary draft from machine generated content. The OneClick Dictionary control screen in Sketch Engine is shown in Figure 1.



D6.1 Recommendations on legal and IPR issues for lexicography

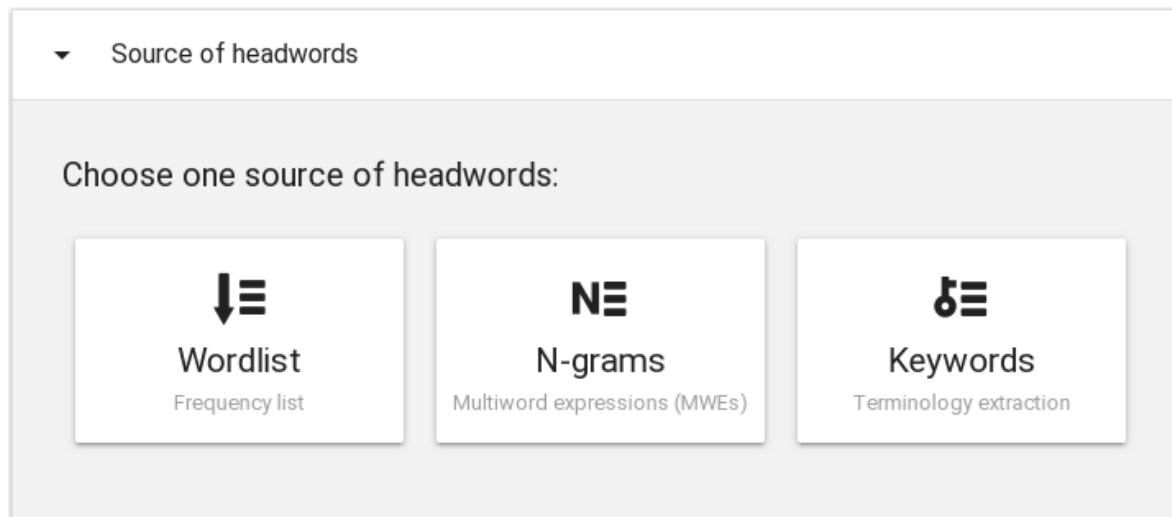


Figure 1: OneClick Dictionary – the dictionary drafting module in Sketch Engine.

The user can select which parts of the dictionary entry should be generated. The following options are available:

headword list

The headword list is the pivotal element of any dictionary. It defines the nature of the dictionary - the size of the dictionary but also how general or specialized the dictionary will be. The content of the headword list is affected by the source data, the corpus, and also by the technology used for its generation. OneClick Dictionary provides three distinct options of generating the headword list:

- a plain wordlist according to users criteria, such as all lemma with a certain minimum frequency
- a wordlist of n-grams, such as bigrams of all nouns
- a wordlist of single- and multi-word keywords in the corpus

headwords from the wordlist tool - The wordlist tool calculates the frequency of each lemma, which is the base form of of a word, in the whole corpus and produces a frequency list. The user defines the dictionary size by specifying how many top frequent lemmas should be used as headwords. This option is mainly intended for building large lexicographic works such as major national dictionaries providing information about how the language is used in general. Typically, the largest corpus in the language will be used as data source.

D6.1 Recommendations on legal and IPR issues for lexicography

headwords from the n-grams tools - The n-gram tools calculates all n-grams in the corpus up to a certain length (6 by default). Such n-grams represent multi-word expression candidates for dictionary inclusion and are subject to the same criteria like a plain word list: settings of minimum and maximum frequency or filtering by a regular expression.

headwords from the Keywords & Terms tool - This option is best suited for creating specialized or domain specific lexicographic works such as specialized dictionaries and also glossaries and other types works based on terminology lists. A specialized corpus would normally be used as data source. If such a corpus is not available, Sketch Engine features a built-in tool for creating domain specific corpora by automatically identifying and downloading relevant content from the web. The downloaded content is automatically processed into a full-fledged corpus in Sketch Engine ready to be used with the Keywords & Terms tool. This tool will automatically identify terminology, i.e. words and phrases which are typical of the corpus or which represent the topic of the corpus.

The resulting headword list can be optionally supplemented with any of the following types of information:

frequency

Based on the statistics extracted from the corpus, each headword can be supplemented with frequency information which will help the user understand whether the word is a common, frequently-used word or whether the word is used only rarely.

collocations

Selecting collocations will trigger the word sketch tool which analyses the context in which each of the headwords appear and identifies words which form the most typical combinations. Collocations are valuable additions to dictionary entries because they help the user use the headword correctly. Collocations also help understand the different contexts in which the headword is typically used.



example**sentences**

This option makes use of the concordance and the GDEX technology. GDEX stands for “Good Dictionary EXamples”. It is a system for the evaluation of sentences with respect to their suitability to serve as dictionary examples or good examples for teaching purposes. Sentences are evaluated with respect to their length, use of complicated vocabulary, presence of controversial topics (politics, religion...), sufficient context, references pointing outside of the sentence (e.g. pronouns), brand names and other criteria. The sentences with the highest GDEX scores are added to the automatically generated dictionary entries.

translations

If a parallel multilingual corpus is available for the languages, translations can also be generated automatically from the corpus data. Typically, more than one translation candidate will be added to the dictionary entry for the editor to decide which should be included in the final dictionary.

synonyms

Synonyms are generated with the help of the thesaurus in Sketch Engine. The thesaurus tool analyses the corpus and uses statistics motivated by the principles of distributional semantics to identify words which appear in similar contexts. Such words are also similar in meaning. A very large dataset is required for the thesaurus to provide high-quality results. A number of synonyms is generated for each dictionary entry for the editor to decide on the ones that should be included.

word forms

For each headword, a list of all existing word forms can be generated. This is useful especially with morphologically rich languages where the user might need to check what the correct word forms are.



part of speech

If the source corpus is tagged for part of speech (POS), the POS label can be added to the headword automatically. Sketch Engine currently supports automatic POS tagging in over 40 languages.

word sense disambiguation

The user may optionally request that word senses should be identified automatically (Figure 2). With this option active, each dictionary entry will be automatically subdivided into a number of word senses. The dictionary entry parts outlined above will be generated for each sense separately. The editor can then edit or combine the senses identified automatically or introduce new senses manually.



Figure 2: A dictionary entry with word senses disambiguated automatically by Sketch Engine.

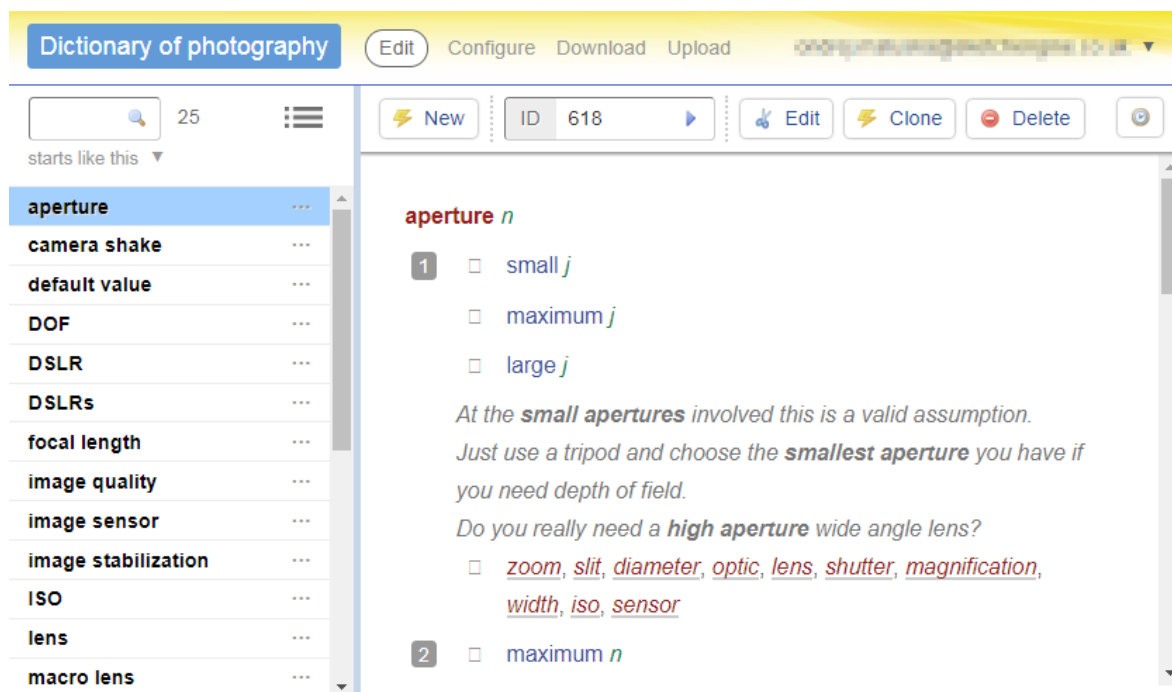
The automatically generated dictionary draft is pushed into Lemony for editing and publishing online.

4 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. Lexonomy is designed to interact with Sketch Engine and its corpora.



The screenshot shows the Lexonomy editor interface. At the top, there is a navigation bar with 'Dictionary of photography' and buttons for 'Edit', 'Configure', 'Download', and 'Upload'. Below this, there is a search bar with '25' results and a 'starts like this' dropdown. A sidebar on the left lists various terms: aperture, camera shake, default value, DOF, DSLR, DSLRs, focal length, image quality, image sensor, image stabilization, ISO, lens, and macro lens. The main content area displays the entry for 'aperture n'. It includes a list of sub-entries: '1' with checkboxes for 'small j', 'maximum j', and 'large j'; a paragraph of text: '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?'; and '2' with a checkbox for 'maximum n'. The text contains several terms in red, likely indicating links or specific annotations.

Figure 3: A dictionary entry in the Lexonomy editor.

4.1 Push model

The push model refers to the initial dictionary draft generation. The process starts in Sketch Engine and requires that the user selects the corpus that should be used as the source data for the dictionary. Then the user decides how the dictionary headword list should be generated. Whether the dictionary headwords should be selected based on frequency using the wordlist tool or whether the headwords should be selected from the terminology contained in the corpus using the Keywords & Terms tool. Then the user configures which parts of the dictionary entry should be generated (collocations, example sentences, synonyms, frequency information etc.). Sketch Engine then analyses the corpus and generates the required number of dictionary headwords with the required content and pushes, or exports, the automatically generated dictionary draft into Lexonomy where it is ready for further editing and for publication online.

4.2 Pull model

The pull model is associated with the process of post-editing the dictionary draft in Lexonomy. When the user, the dictionary editor, works with the automatically generated content, it might become necessary to check the source corpus or it might be necessary to generate additional information for the dictionary entry, for example, more collocations might be needed or different example sentences might be required. This is when the pull model comes in. Lexonomy is designed to communicate with Sketch Engine. A dictionary in Lexonomy can be linked to a specific corpus in Sketch Engine so that additional data can be pulled from the corpus if needed.

4.3 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



D6.1 Recommendations on legal and IPR issues for lexicography

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 are 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.

4.4 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 6) 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.

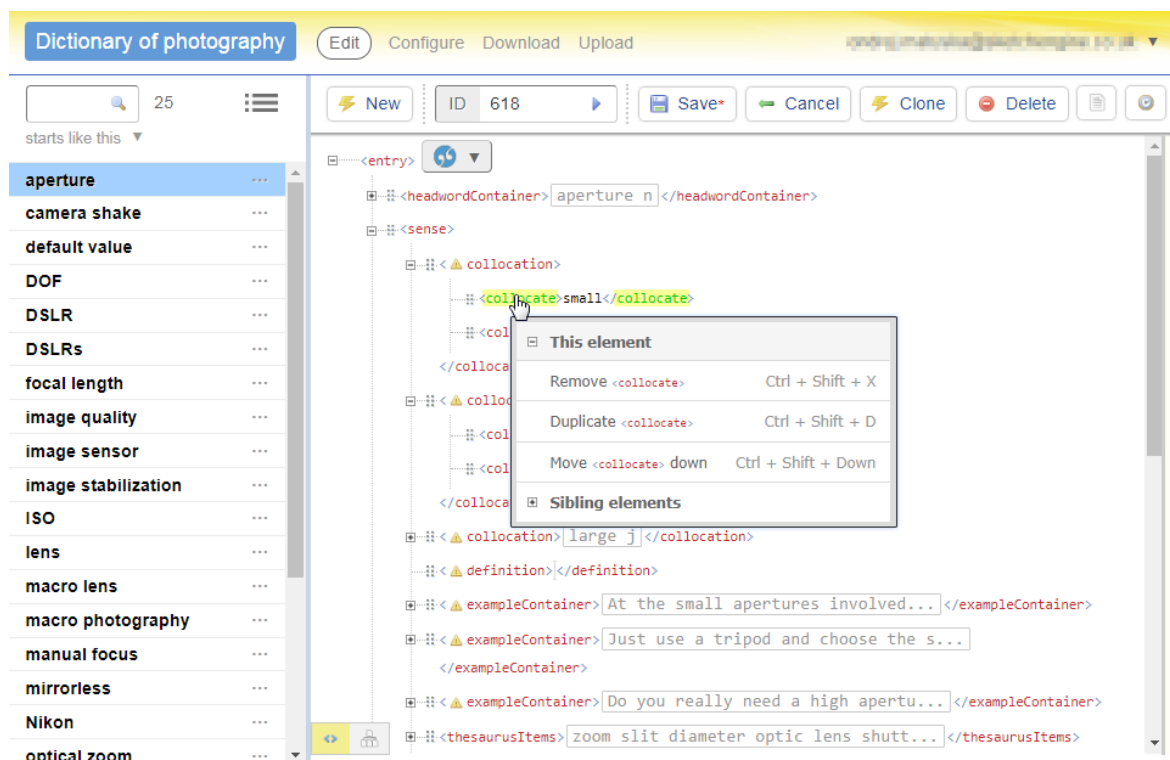


Figure 4: Editing particular attributes of a dictionary entry within Lexonomy.



D6.1 Recommendations on legal and IPR issues for lexicography

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

4.5 Dual editing interface

The editing interface now allows the user to switch from an XML-based to a more visual layout suitable for less IT-aware or less technical users. (see Figure 5)



Figure 5: Dictionary entry editor - visual layout.

5 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.

