elexis european lexicographic infrastructure

D8.2

Periodic assessment of LEX1, LEX2 and LEX3 – second report

Author(s): Miloš Jakubíček, Ondřej Matuška, Michal Cukr, Simon Krek, Michael Rundell

Date: 31. 1. 2020

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D8.2 Periodic assessment of LEX1, LEX2 and LEX3 – second report

Deliverable Number:	D8.2
Dissemination Level:	Public
Delivery Date:	31. 1. 2020
Version:	1.0
Author(s):	Miloš Jakubíček, Ondřej Matuška, Michal Cukr, Simon Krek, Michael Rundell

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
15. 1. 2020	Intial draft	Matuška
20. 1. 2020	Added technical provisions	Jakubíček
27. 1. 2020	Assessment by AB member	Rundell

Table of Contents

List of Figures

1 Introduction

This report provides an assessment of the three parts of the ELEXIS infrastructure (LEX1, LEX2 and LEX3) during the second year of the project (M12-M24). Each of these parts is represented with its own task under the WP8 and this report is structured accordingly.

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 Part 1: Lexicographic data integration

The central LEX1/LEX2/(LEX3) infrastructure will be launched in March 2020 and consists of six interlinked service modules which draw on tools alreday developed within LEX2. The modules, shown in Figure 1, are:

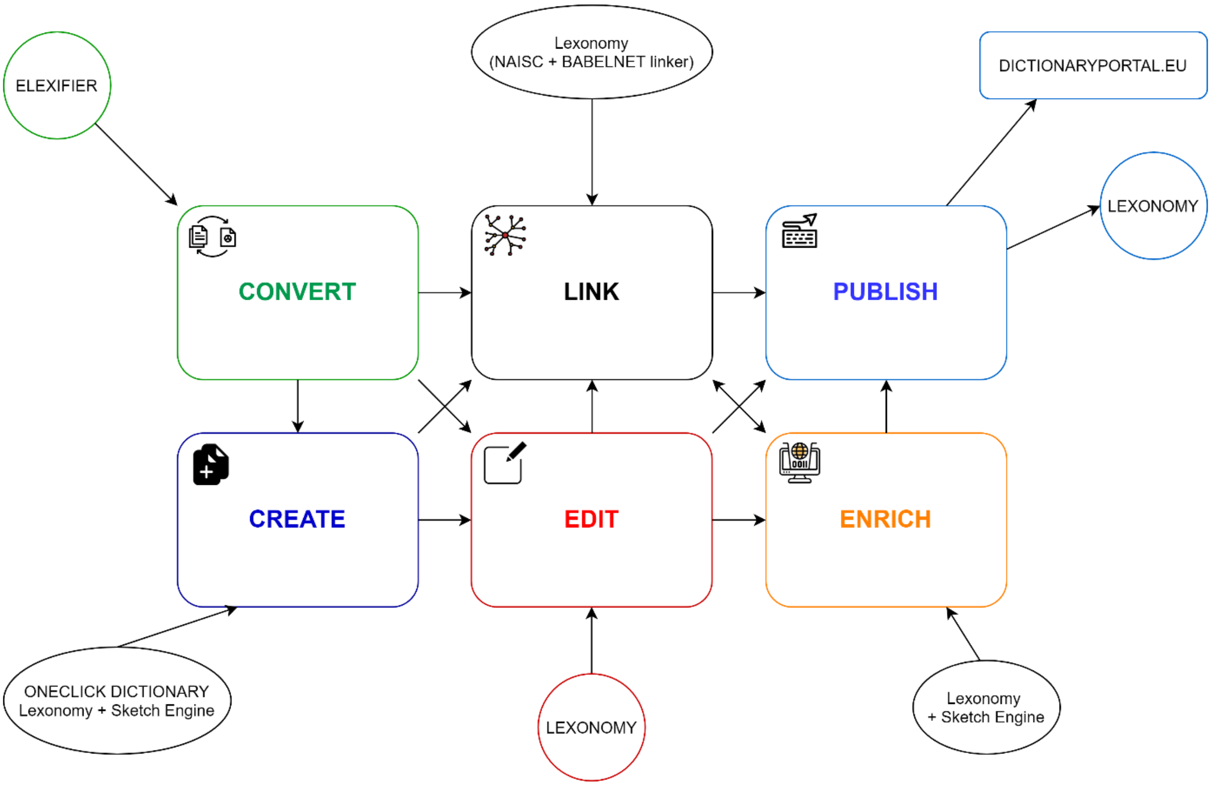


Figure 1: LEX 1 modules.

- CONVERT module provides conversion of dictionaries in various formats into TEI-Lex0 or Ontolex-Lemon format, using the Elexifier tool. [ALREADY AVAILABLE AS PART OF LEX1]
- CREATE module enables the creation of a completely new dictionary, also with the use of OneClick Dictionary (LEX2) that draws on the link between the Sketch Engine corpus tool and Lexonomy dictionary writing system. [ALREADY AVAILABLE AS PART OF LEX2]

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.

- EDIT module provides the option of uploading and compilation of the dictionary in the Lexonomy dictionary writing system. [ALREADY AVAILABLE AS PART OF LEX1]
- ENRICH module enables the addition of new types of data such as corpus examples, images, and audio and video recordings to dictionaries. This can be done on existing dictionaries imported into Lexonomy, or newly created content. [WILL BE AVAILABLE IN M36]
- LINK module provides services of linking dictionary entries or parts of entries with other monolingual and multilngual lexicographic resources. [WILL BE AVAILABLE IN M36]
- PUBLISH module support publication of dictionary content which can be done in two ways: using the publishing function in the Lexonomy dictionary writing system, or via dictionaryportal.eu website. [LEXINOMY IS AVAILABLE; DICTIONARY PORTAL WILL BE AVAILABLE IN M36]

The LEX1 infrastructure is hosted and operated by Jozef Stefan Institute.

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 Part 2: Platform for dictionary drafting from corpora (LEX2)

The task (T8.2) relates to this part of the infrastructure. It was launched in M6. The infrastructure referred to as LEX 2 is run by Lexical Computing and hosted on their servers. This report outlines:

1. tools and services that are part of the infrastructure
2. technical provisions that were taken to operate the infrastructure and provide access to it
3. statistics of the usage of the infrastructure

3.1 Tools and services in LEX2

3.1.1 SketchEngine

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 or language teachers, language learners and others in need of understanding how language is used. Sketch Engine currently contains corpora in almost 100 languages and allows users to upload texts in any of those languages to build their own corpora. 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 corpora in the language available.

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

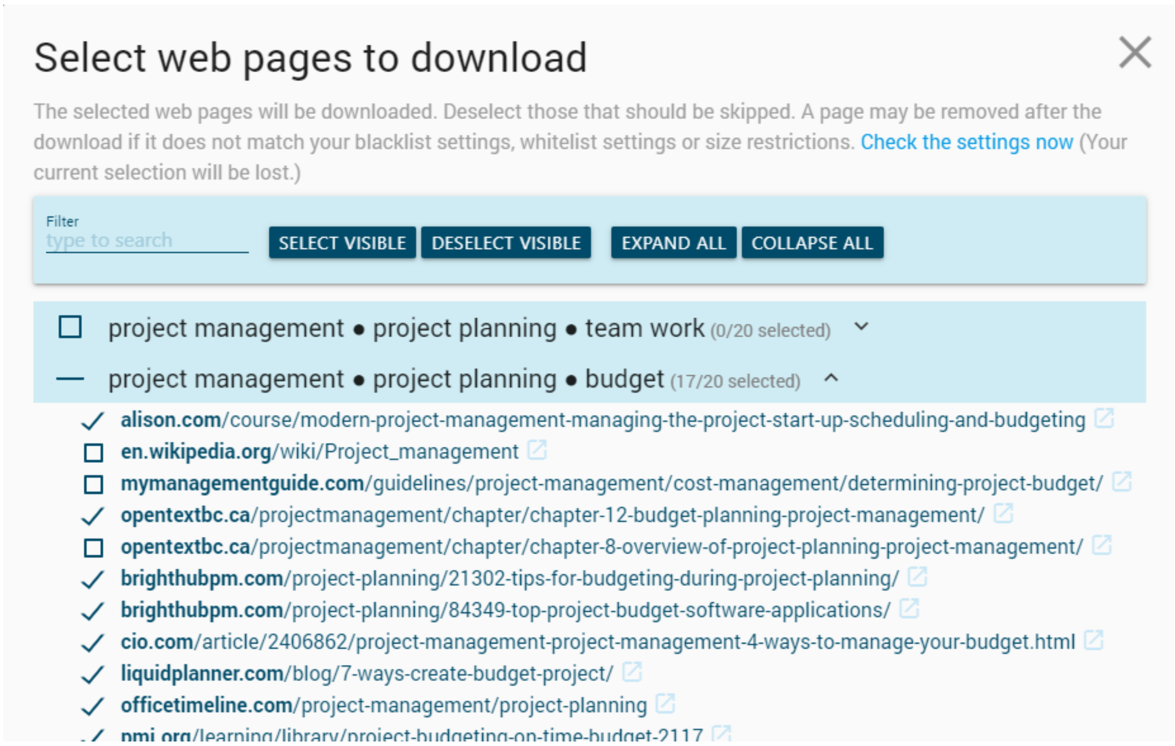
The tools in Sketch Engine include:

Corpus building tool – This fully automatic tool is designed for building corpora of any size including instant building of small and highly-specialized corpora for immediate use. It was developed with a non-IT user in mind and completely automates tasks such as tokenization, part-of-speech tagging and lemmatization. Applying such tools manually requires programmatic and IT expertise which may represent an insurmountable obstacle for many

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.

researchers attempting corpus-oriented methods. Sketch Engine makes corpus building accessible to anybody with only general computer skills.

To provide even better control of the content of user corpora automatically built from web material, management of the pages to download has been added as shown in Figure 2. The user can review the content identified by Sketch Engine and manually deselect groups or individual sources to remove undesirable content. This will produce a cleaner and more focussed data resource.



Select web pages to download

The selected web pages will be downloaded. Deselect those that should be skipped. A page may be removed after the download if it does not match your blacklist settings, whitelist settings or size restrictions. Check the settings now (Your current selection will be lost.)

☐ project management • project planning • team work (0/20 selected)

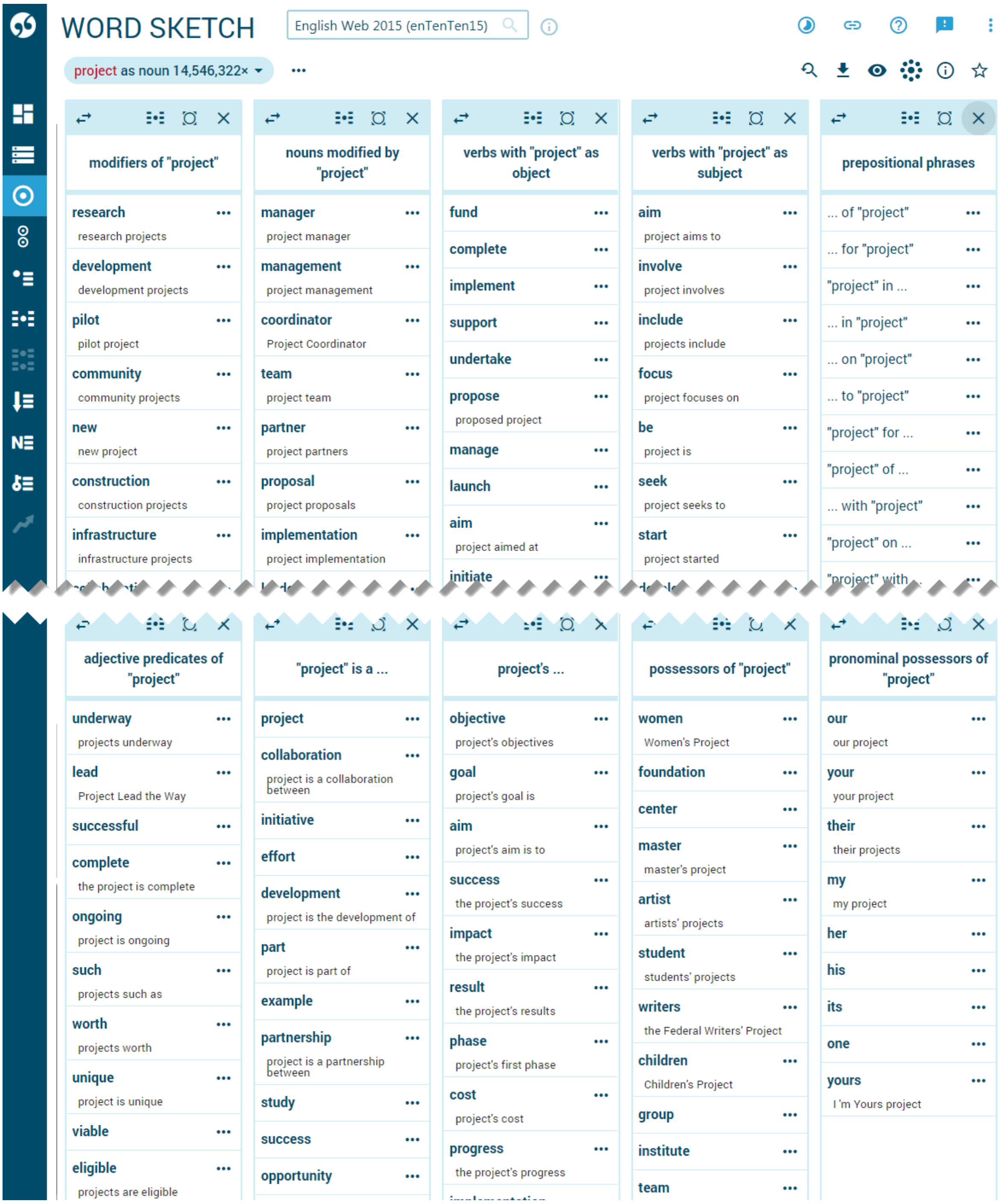
— project management • project planning • budget (17/20 selected)

- ✓ alison.com/course/modern-project-management-managing-the-project-start-up-scheduling-and-budgeting
- ☐ en.wikipedia.org/wiki/Project_management
- ☐ mymanagementguide.com/guidelines/project-management/cost-management/determining-project-budget/
- ✓ opentextbc.ca/projectmanagement/chapter/chapter-12-budget-planning-project-management/
- ☐ opentextbc.ca/projectmanagement/chapter/chapter-8-overview-of-project-planning-project-management/
- ✓ brighthubpm.com/project-planning/21302-tips-for-budgeting-during-project-planning/
- ✓ brighthubpm.com/project-planning/84349-top-project-budget-software-applications/
- ✓ cio.com/article/2406862/project-management-project-management-4-ways-to-manage-your-budget.html
- ✓ liquidplanner.com/blog/7-ways-create-budget-project/
- ✓ officetimeline.com/project-management/project-planning
- ✓ pmi.org/learning/library/project-budgeting-on-time-budget-2117

Figure 2: An interface allowing the user to approve the web resources identified by Sketch Engine.

Word Sketch – The key tool in Sketch Engine which gave the system its name is designed to quickly identify all occurrences of a word in a corpus of any size, process the contexts in which the word appears and display the words (collocates) which typically appear together with the search word and form the collocations. The collocates are presented to the user in the form of a table and categorized into dozens of categories by the type of grammatical relation they have with the search word. An illustrative example of a word sketch results is represented by Figure 3.

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.



WORD SKETCH

English Web 2015 (enTenTen15)

project as noun 14,546,322×

modifiers of "project"
- research — research projects
- development — development projects
- pilot — pilot project
- community — community projects
- new — new project
- construction — construction projects
- infrastructure — infrastructure projects

nouns modified by "project"
- manager — project manager
- management — project management
- coordinator — Project Coordinator
- team — project team
- partner — project partners
- proposal — project proposals
- implementation — project implementation

verbs with "project" as object
- fund
- complete
- implement
- support
- undertake
- propose — proposed project
- manage
- launch
- aim — project aimed at
- initiate

verbs with "project" as subject
- aim — project aims to
- involve — project involves
- include — projects include
- focus — project focuses on
- be — project is
- seek — project seeks to
- start — project started

prepositional phrases
- ... of "project"
- ... for "project"
- "project" in ...
- ... in "project"
- ... on "project"
- ... to "project"
- "project" for ...
- "project" of ...
- ... with "project"
- "project" on ...
- "project" with ...

adjective predicates of "project"
- underway — projects underway
- lead — Project Lead the Way
- successful
- complete — the project is complete
- ongoing — project is ongoing
- such — projects such as
- worth — projects worth
- unique — project is unique
- viable
- eligible — projects are eligible

"project" is a ...
- project
- collaboration — project is a collaboration between
- initiative
- effort
- development — project is the development of
- part — project is part of
- example
- partnership — project is a partnership between
- study
- success
- opportunity

project's ...
- objective — project's objectives
- goal — project's goal is
- aim — project's aim is to
- success — the project's success
- impact — the project's impact
- result — the project's results
- phase — project's first phase
- cost — project's cost
- progress — the project's progress

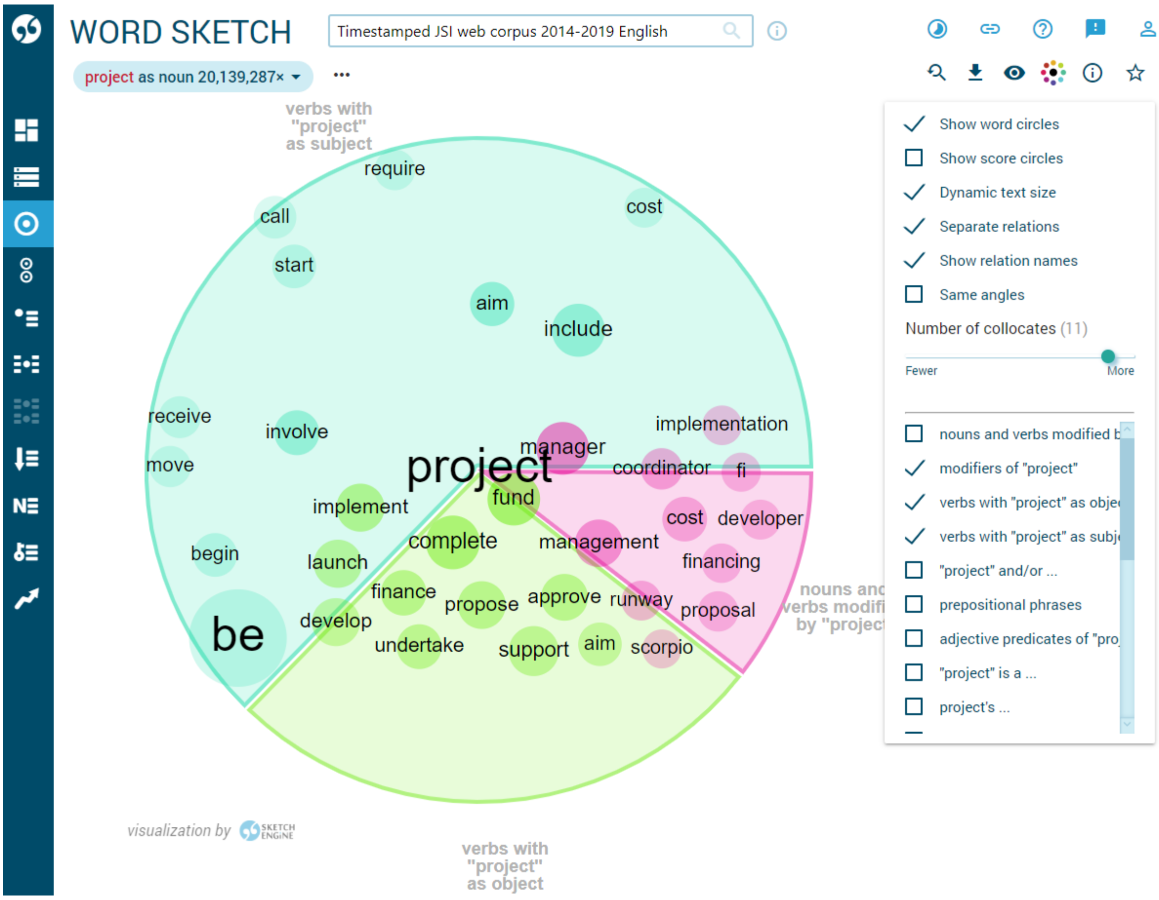
possessors of "project"
- women — Women's Project
- foundation
- center
- master — master's project
- artist — artists' projects
- student — students' projects
- writers — the Federal Writers' Project
- children — Children's Project
- group
- institute
- team

pronominal possessors of "project"
- our — our project
- your — your project
- their — their projects
- my — my project
- her
- his
- its
- one
- yours — I 'm Yours project

Figure 3: Results for the noun "project" from the Word sketch feature of Sketch Engine.

This visualisation (Figure 4) of the data promotes quick understanding in which contexts it typically appears and how it is used. This is extremely effective because it negates the need to study, referring to the screenshot now, all the 20 million of occurrences of the word *project*.

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.



✓ Show word circles
□ Show score circles
✓ Dynamic text size
✓ Separate relations
✓ Show relation names
□ Same angles

Number of collocates (11)

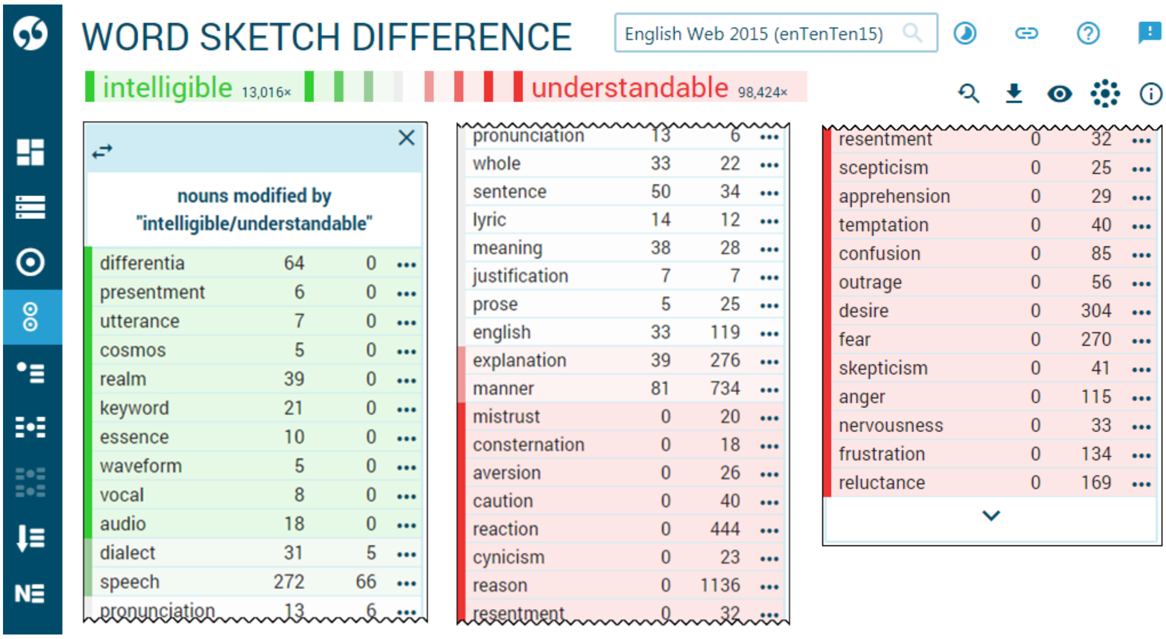
Fewer More

□ nouns and verbs modified b
✓ modifiers of "project"
✓ verbs with "project" as obje
✓ verbs with "project" as subj
□ "project" and/or ...
□ prepositional phrases
□ adjective predicates of "pro
□ "project" is a ...
□ project's ...

Figure 4: Results for the noun "project" from the Word sketch feature of Sketch Engine.

Word Sketch Difference – An extension of the word sketch tool used to compare the use and meaning of two words via the collocations they form. Each word is assigned a colour and the collocates carry the shade of the colour based on how strongly they are related to one word or the other. Figure 5 shows the layout of the tool. Find more information in [1]

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.



WORD SKETCH DIFFERENCE

English Web 2015 (enTenTen15)

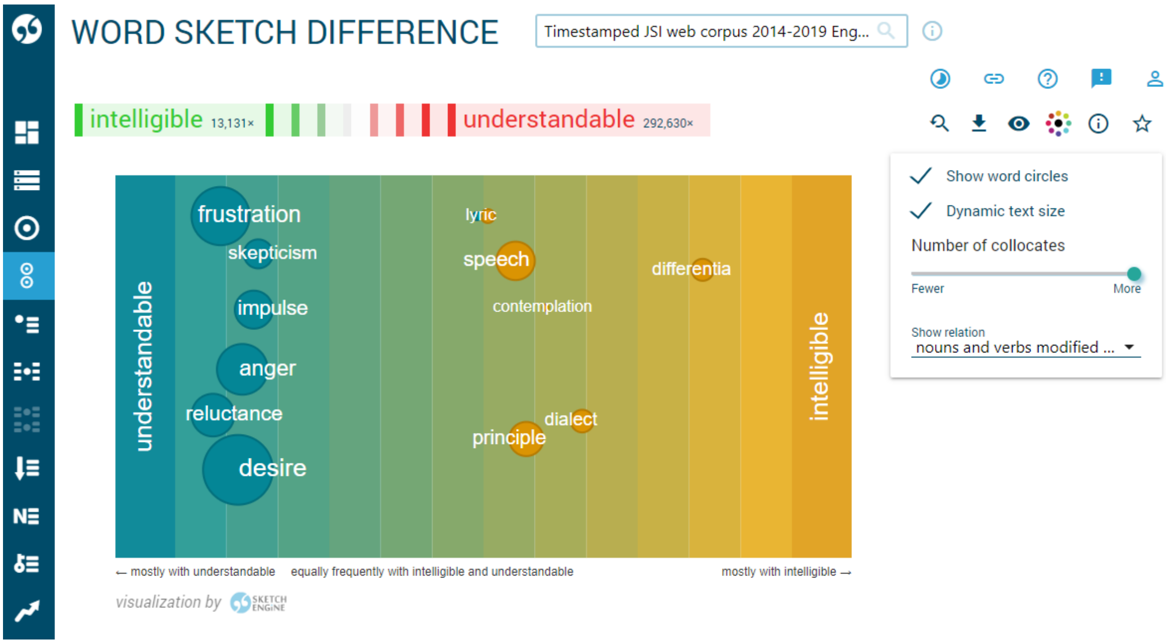
intelligible 13,016× understandable 98,424×

nouns modified by "intelligible/understandable"

differentia	64	0	•••
presentment	6	0	•••
utterance	7	0	•••
cosmos	5	0	•••
realm	39	0	•••
keyword	21	0	•••
essence	10	0	•••
waveform	5	0	•••
vocal	8	0	•••
audio	18	0	•••
dialect	31	5	•••
speech	272	66	•••
pronunciation	13	6	•••
whole	33	22	•••
sentence	50	34	•••
lyric	14	12	•••
meaning	38	28	•••
justification	7	7	•••
prose	5	25	•••
english	33	119	•••
explanation	39	276	•••
manner	81	734	•••
mistrust	0	20	•••
consternation	0	18	•••
aversion	0	26	•••
caution	0	40	•••
reaction	0	444	•••
cynicism	0	23	•••
reason	0	1136	•••
resentment	0	32	•••
scepticism	0	25	•••
apprehension	0	29	•••
temptation	0	40	•••
confusion	0	85	•••
outrage	0	56	•••
desire	0	304	•••
fear	0	270	•••
skepticism	0	41	•••
anger	0	115	•••
nervousness	0	33	•••
frustration	0	134	•••
reluctance	0	169	•••

Figure 5: A layout of the word sketch difference tool for the adverbs “intelligible” and “understandable”.

Visualisations (Figure 6) have been developed to facilitate the understanding of the data generated by the word sketch difference.



WORD SKETCH DIFFERENCE

Timestamped JSI web corpus 2014-2019 Eng...

intelligible 13,131× understandable 292,630×

✓ Show word circles

✓ Dynamic text size

Number of collocates: More

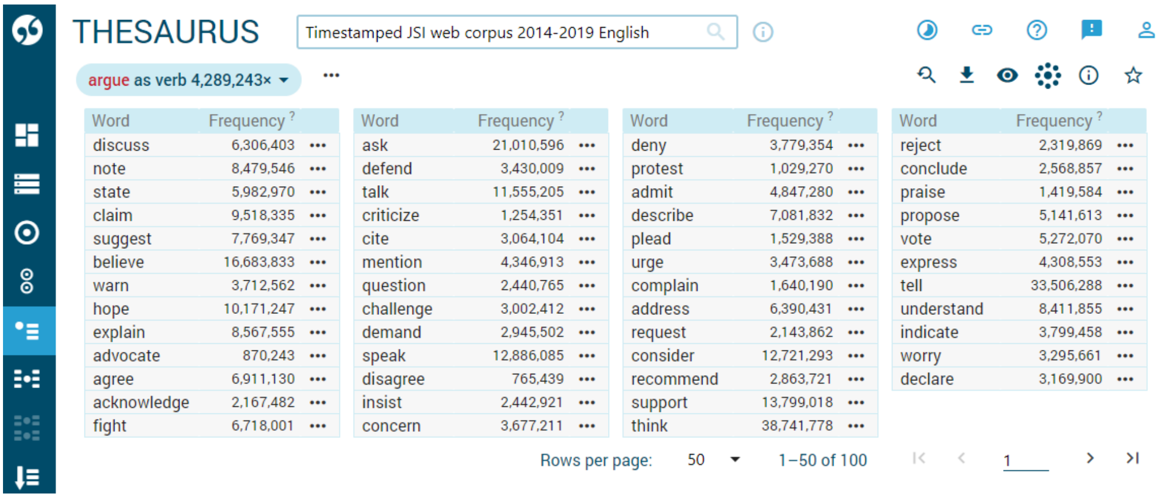
Show relation: nouns and verbs modified ...

understandable
frustration
skepticism
impulse
anger
reluctance
desire
lyric
speech
contemplation
principle
dialect
differentia
intelligible
Fewer
More
← mostly with understandable
equally frequently with intelligible and understandable
mostly with intelligible →
visualization by SKETCH ENGINE

Figure 6: Visualisation of the differences in collocations modifying the words "intelligible" and "understandable".

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.

Thesaurus – This tool draws on the theory of distributional semantics to automatically generate a thesaurus – a list of synonyms or words belonging to the same semantic category. Unlike man-made thesauri, the automatically generated thesaurus in Sketch Engine can be generated for any word in the language, provided a sufficient number of occurrences is found in the corpus. The thesaurus is important for lexicography, language teaching and also NLP and IT applications.[1]



THESAURUS

Timestamped JSI web corpus 2014-2019 English

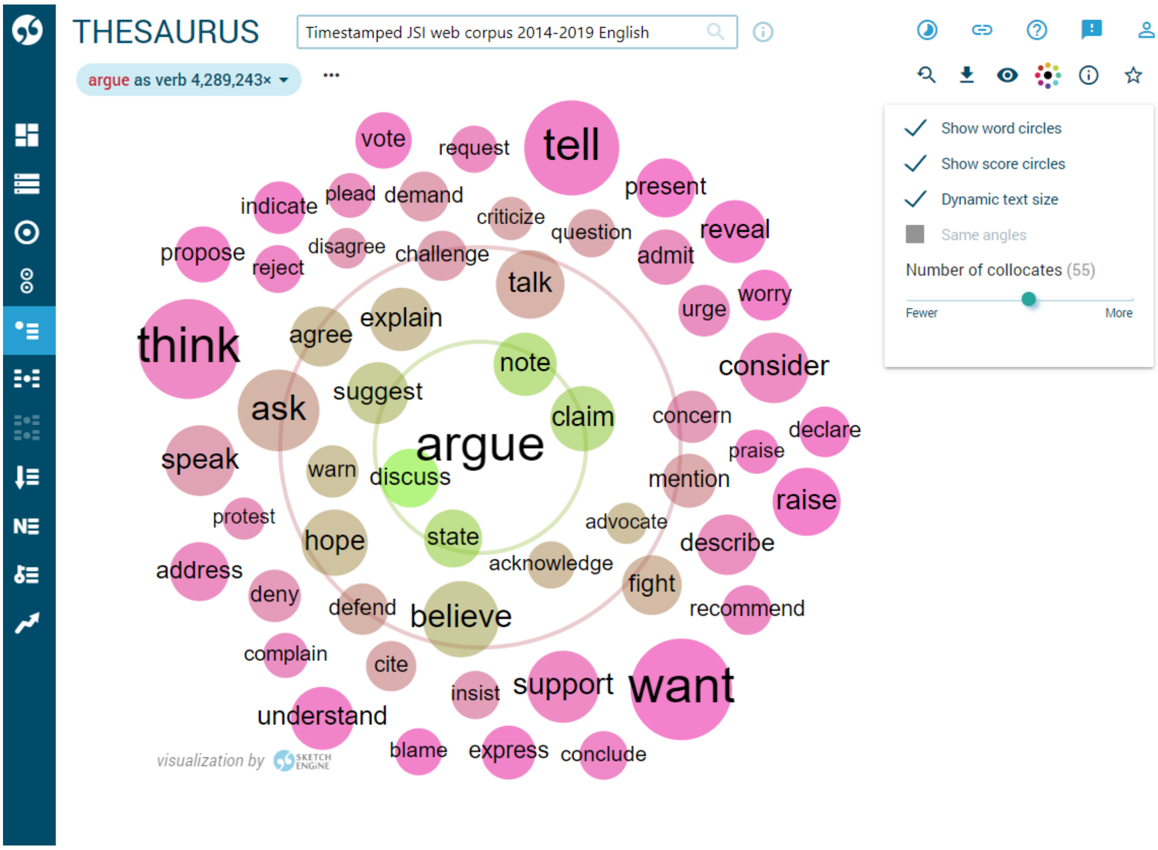
argue as verb 4,289,243×

Word	Frequency ?
discuss	6,306,403
note	8,479,546
state	5,982,970
claim	9,518,335
suggest	7,769,347
believe	16,683,833
warn	3,712,562
hope	10,171,247
explain	8,567,555
advocate	870,243
agree	6,911,130
acknowledge	2,167,482
fight	6,718,001
ask	21,010,596
defend	3,430,009
talk	11,555,205
criticize	1,254,351
cite	3,064,104
mention	4,346,913
question	2,440,765
challenge	3,002,412
demand	2,945,502
speak	12,886,085
disagree	765,439
insist	2,442,921
concern	3,677,211
deny	3,779,354
protest	1,029,270
admit	4,847,280
describe	7,081,832
plead	1,529,388
urge	3,473,688
complain	1,640,190
address	6,390,431
request	2,143,862
consider	12,721,293
recommend	2,863,721
support	13,799,018
think	38,741,778
reject	2,319,869
conclude	2,568,857
praise	1,419,584
propose	5,141,613
vote	5,272,070
express	4,308,553
tell	33,506,288
understand	8,411,855
indicate	3,799,458
worry	3,295,661
declare	3,169,900

Figure 7: A thesaurus list for the word "argue" with words sorted by the similarity to argue based on the principles of distributional semantics.

Visualisations (Figure 8) have been developed to facilitate the understanding of the data. The distance from the centre indicates the similarity, the size of the circle corresponds to frequency.

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.



THESAURUS

Timestamped JSI web corpus 2014-2019 English

argue as verb 4,289,243×

✓ Show word circles

✓ Show score circles

✓ Dynamic text size

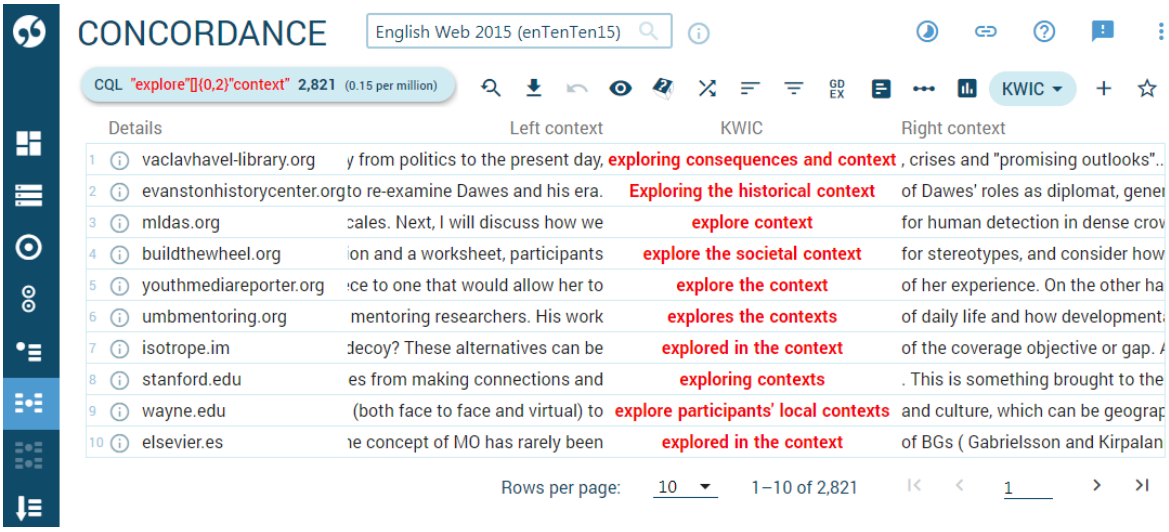
□ Same angles

Number of collocates (55)

Figure 8: A visualisation of the thesaurus for the word "argue".

Concordance – The source data (sentences) from which results in any tool were generated can be seen in the concordance tool. The user always has direct access to the concrete examples of the words or phrases in context to drill down in more detail. The concordance can also be used on its own to search for examples of words or phrases (Figure 8) and also of lexical and grammatical structures without specifying concrete words. A multitude of search options is available and a whole suite of result processing tools is available such as filtering, sorting, calculating frequencies.[1] The view options allow the visualisation of various additional information about the texts studied.

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.



CONCORDANCE English Web 2015 (enTenTen15)

CQL "explore"[]{0,2}"context" 2,821 (0.15 per million)

	Details	Left context	KWIC	Right context
1	vaclavhavel-library.org	y from politics to the present day,	exploring consequences and context	, crises and "promising outlooks"..
2	evanstonhistorycenter.org	to re-examine Dawes and his era.	Exploring the historical context	of Dawes' roles as diplomat, gener
3	mldas.org	cales. Next, I will discuss how we	explore context	for human detection in dense crov
4	buildthewheel.org	ion and a worksheet, participants	explore the societal context	for stereotypes, and consider how
5	youthmediareporter.org	ce to one that would allow her to	explore the context	of her experience. On the other ha
6	umbmentoring.org	mentoring researchers. His work	explores the contexts	of daily life and how development
7	isotrope.im	decoy? These alternatives can be	explored in the context	of the coverage objective or gap.
8	stanford.edu	es from making connections and	exploring contexts	. This is something brought to the
9	wayne.edu	(both face to face and virtual) to	explore participants' local contexts	and culture, which can be geograp
10	elsevier.es	ie concept of MO has rarely been	explored in the context	of BGs (Gabrielsson and Kirpalan

Rows per page: 10 1–10 of 2,821 1

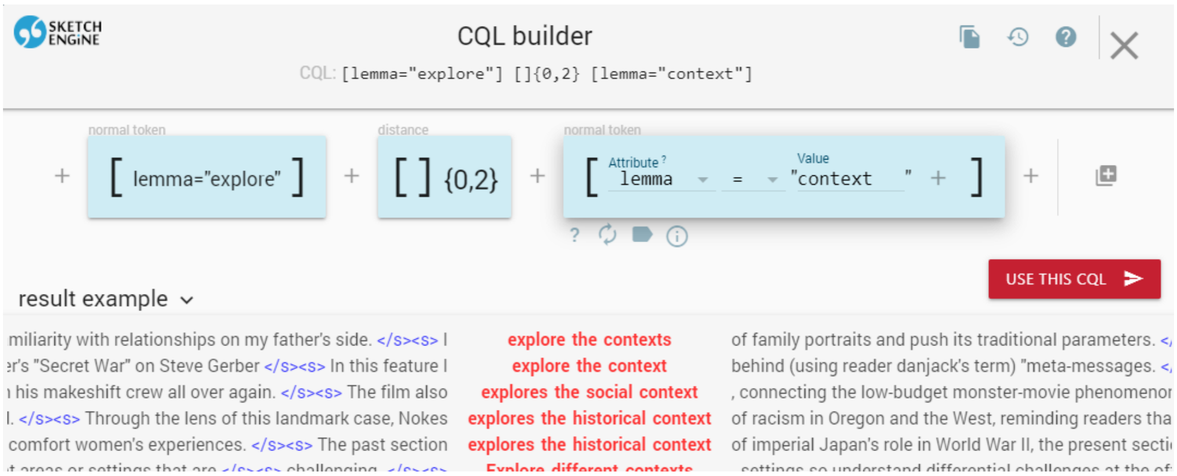
Figure 9: A concordance showing examples of a multiword query.

The concordance also features a unique GDEX technology which, when activated, evaluates the sentences with respect to their suitability to serve as Good Dictionary EXamples [2]. This functionality is extremely useful for lexicography but also for the development of language teaching materials and in language teaching in general.

The concordance now allows for more granular frequency analysis through the use of labels. The labels may now be assigned to words within a multiword query and the frequency analysis can then be restricted to only the labelled positions.

To facilitate the construction of complex queries using the Corpus Query Language, a new CQL builder had been developed. The CQL builder is an important improvement for non-IT and less technical users who can now construct complex queries correctly and easily by selecting query elements from predefined options rather than being forced to type all the required special characters.

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.



CQL: [lemma="explore"] []{0,2} [lemma="context"]

[lemma="explore"] + [ ] {0,2} + [lemma = "context " +]

```
miliarity with relationships on my father's side. </s><s> I        explore the contexts              of family portraits and push its traditional parameters. <
er's "Secret War" on Steve Gerber </s><s> In this feature I         explore the context               behind (using reader danjack's term) "meta-messages. <
his makeshift crew all over again. </s><s> The film also            explores the social context       , connecting the low-budget monster-movie phenomenon
l. </s><s> Through the lens of this landmark case, Nokes            explores the historical context   of racism in Oregon and the West, reminding readers tha
comfort women's experiences. </s><s> The past section               explores the historical context   of imperial Japan's role in World War II, the present secti
t areas or settings that are </s><s> challenging. </s><s>           Explore different contexts        settings so understand differential challenges at the of
```

Figure 10: The CQL builders with the query used to generate the concordance in Figure 8.

Parallel concordance – It is an extension of the concordance designed to work with multilingual aligned corpora which contain source documents and their translations into another language. [1] They are used in bilingual lexicography to identify translation alternatives and the parallel data themselves are used to generate bilingual and multilingual databases (dictionaries) of translations.

Wordlist – The wordlist tool generates frequency lists of any information in the corpus including metadata. The most frequent uses include the frequency list of all words in the language, all lemmas (based forms or dictionary forms), frequency lists of nouns, verbs, adjectives and other parts of speech or frequency lists of words starting, containing or ending with a particular group of letters. Such lists are used in lexicography, IT and NLP applications and also for understanding the content of the corpus.

N-grams – N-grams are sequences of tokens (or words). In linguistics, they are often referred to as multiword expressions (MWEs). The N-gram tool generates frequency lists of N-grams. The user can define the size of the N-gram (bigram, trigram, up to 6-gram) and also apply various restrictions or filtering similarly as with the wordlist tool. MWEs have its place in

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.

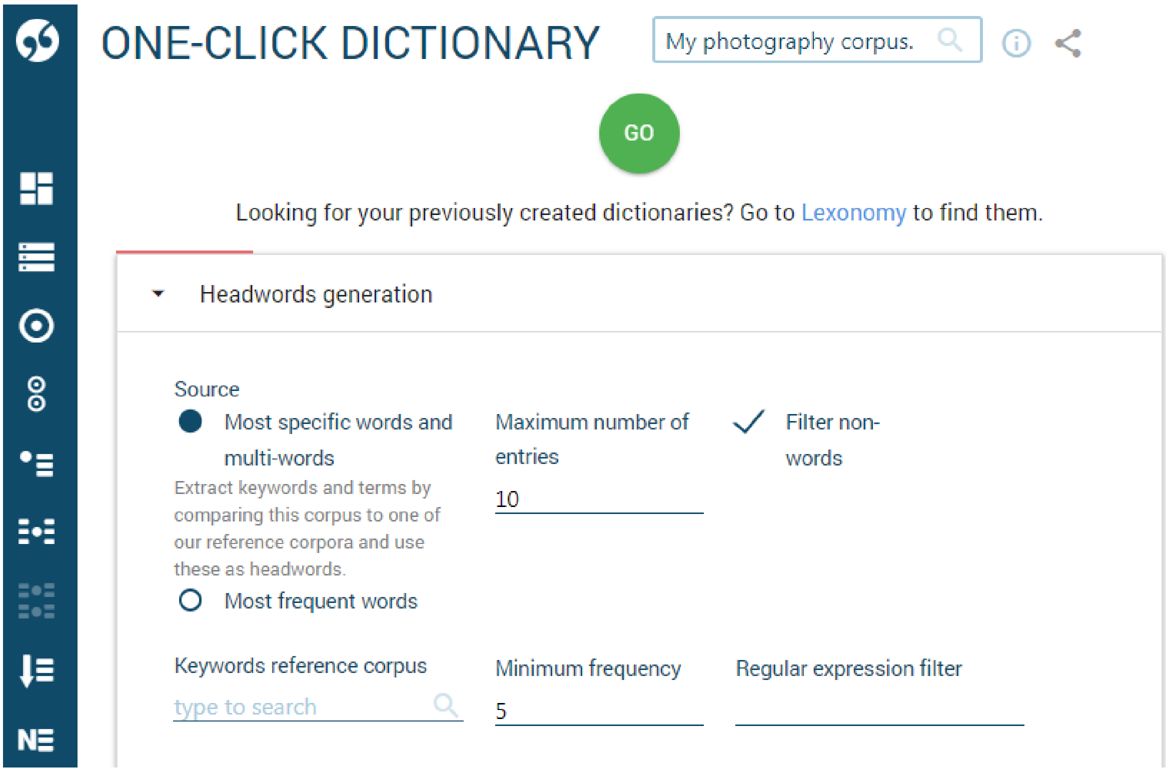
linguistic research and second language acquisition as well as in language modelling for IT and NLP applications.

Keywords & Terms – This tool identifies automatically words and phrases which are typical of the corpus and they define the content or the topic of the corpus. This helps the user check whether the corpus covers a large variety of topics or whether the corpus is a specialized one covering only one or a small selection of related domains. This functionality [1] can also be exploited by translators, interpreters and terminologists for terminology extraction. In addition, it can be used for automatic document classification in IT applications.

Trends – A tool designed to monitor changes in language: new words (neologisms), words going out of use and also words with a sudden peak or slump in use suggesting which topics are becoming more prominent or less talked about. Trends are entirely dependent on time-stamped data which Sketch Engine collects and processes into corpora daily.

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 11). Such dictionary drafts would still need to be post-edited by lexicographers but deleting, amending, rephrasing is more productive than developing dictionary entries from scratch. OneClick Dictionary triggers all the tools described above and produces the 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 11: 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.

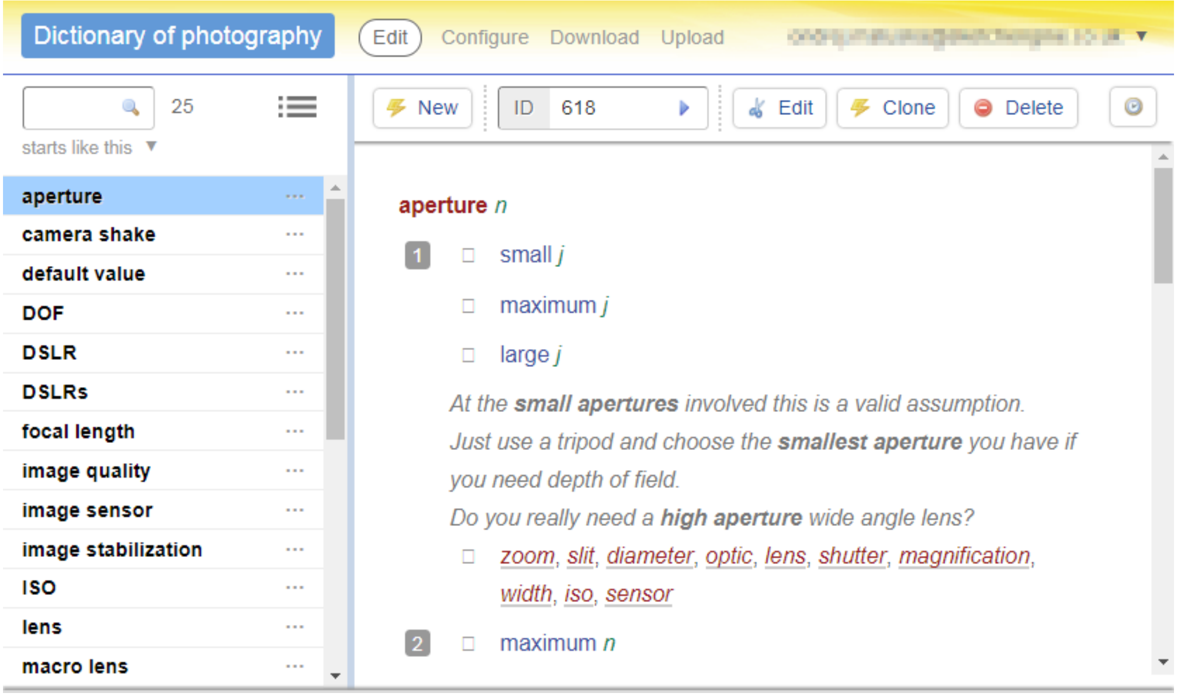
3.1.2 Lexonomy

⟨ LEXONOMY ⟩

access on www.lexonomy.eu

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.

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 user interface.



aperture
camera shake
default value
DOF
DSLR
DSLRs
focal length
image quality
image sensor
image stabilization
ISO
lens
macro lens

aperture *n*

1 □ small *j*

□ maximum *j*

□ large *j*

*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*

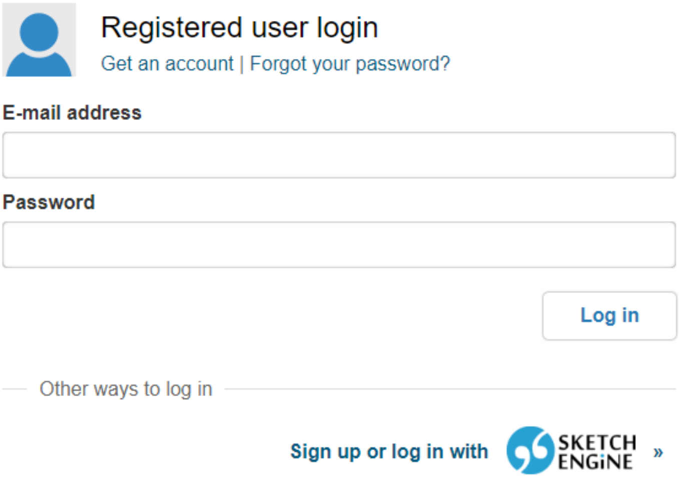
2 □ maximum *n*

Figure 12: A dictionary entry in Lexonomy.

In 2019, the whole system experienced a complete transition from JavaScript to Python. This move was taken to make the system more robust, stable and scalable. The transition also brought about higher speed and better response.

The access to the system has been made easier too. Users no longer have to go through the Lexonomy account registration. They can make use of the "sign in with Sketch Engine" option and use their Sketch Engine login details to access Lexonomy. This option is also compatible with the Single Sign-On for users taking advantage of the ELEXIS-funded access to Sketch Engine (Figure 13).

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.



Registered user login

Get an account | Forgot your password?

E-mail address

Password

Figure 13: Lexonomy sign in via the Sketch Engine account.

3.1.2.1 Push model

Dictionaries in Lexonomy can be created from scratch manually but it is far more effective not to start with an empty dictionary but have a dictionary draft pre-generated from corpus data with the help of tools integrated into Sketch Engine. The data pushed from Sketch Engine into Lexonomy can consist of a plain list of headwords generated based on the word frequency or based on an automatically generated terminology list. The latter is suitable for domain-specific lexicographic works. In addition to this plain list of headwords, the user can decide to push additional information (see the description of OneClick Dictionary above). Pushing the data will create a structured dataset with the respective dictionary entry structure and all the needed elements. The users can make use of predefined dictionary templates but Lexonomy also allows custom dictionary templates which can be used to accommodate specific requirements.

3.1.2.2 Dictionary templates

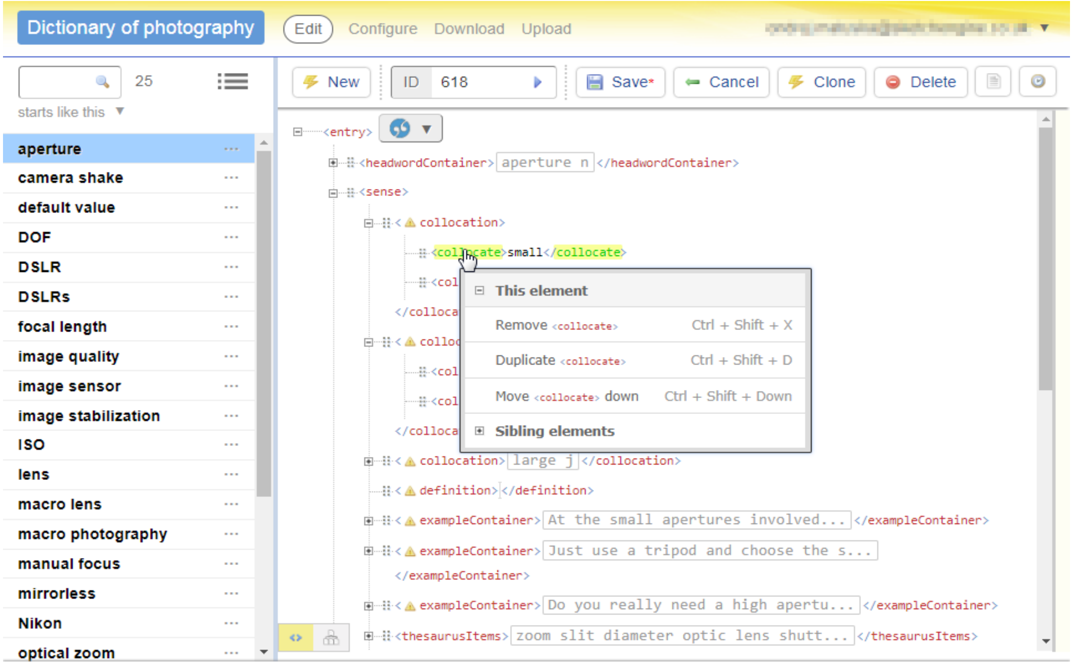
Lexonomy supports dictionary templates which define what elements dictionary entries should or must contain. Each piece of a dictionary entry information such as pronunciation, definition, example, synonym, collocation, translation etc. can be defined as optional or compulsory, the number of such elements 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

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.

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.

3.1.2.3 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 14) and completely eliminates the error-prone procedure of typing the XML code manually.



```
<entry>
  <headwordContainer> aperture n </headwordContainer>
  <sense>
    < collocation>
      <collocate>small</collocate>
      <col
    </colloca
    < colloc
      <col
      <col
    </colloca
    < collocation> large j </collocation>
    < definition></definition>
    < exampleContainer> At the small apertures involved... </exampleContainer>
    < exampleContainer> Just use a tripod and choose the s...
    </exampleContainer>
    < exampleContainer> Do you really need a high apertu... </exampleContainer>
    <thesaurusItems> zoom slit diameter optic lens shutt... </thesaurusItems>
```

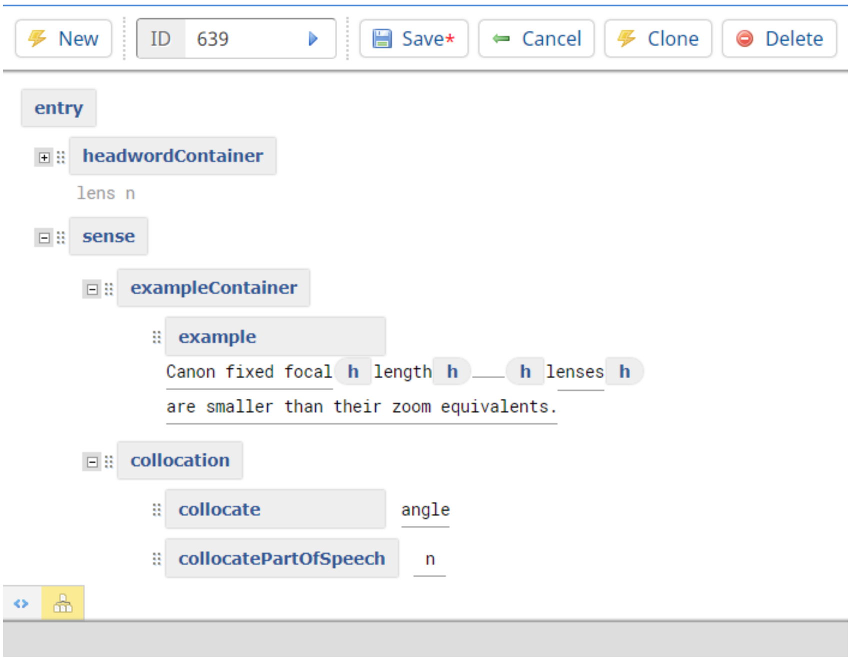
Figure 14: Editing particular attributes of a dictionary entry within Lexonomy.

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

3.1.2.4 Dual editing interface

The editing interface allows 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 15)

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.



lens n

Canon fixed focal h length h h lenses h are smaller than their zoom equivalents.

collocate angle

collocatePartOfSpeech n

Figure 15: Dictionary entry editor - visual layout.

3.1.2.5 User manual

An extended user manual documenting the use of the editing tools in Lexonomy was published on https://www.lexonomy.eu/docs/intro

3.1.2.6 Flags

A system of labelling for dictionary headwords was introduced to facilitate grouping, categorising and classifying dictionary entries as seen in Figure 15. The functionality can be operated with a mouse or via keyboard for greater productivity. The categories can be defined as necessary for each dictionary.

3.1.2.7 Unicode characters

Lexonomy is now capable of handling a large number of writing systems including ones which are not based on the latin script. (Figure 16)

3.1.2.8 Pull model

Whenever the user needs to check the usage in an authentic sample of language, the corpora in Sketch Engine are made accessible directly from the Lexonomy interface as shown in Figure 8. Each dictionary project can be linked to a different corpus in Sketch Engine to acknowledge the fact that a domain-

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.

specific glossary might need to draw data from a different data source (corpus) than a general language dictionary.

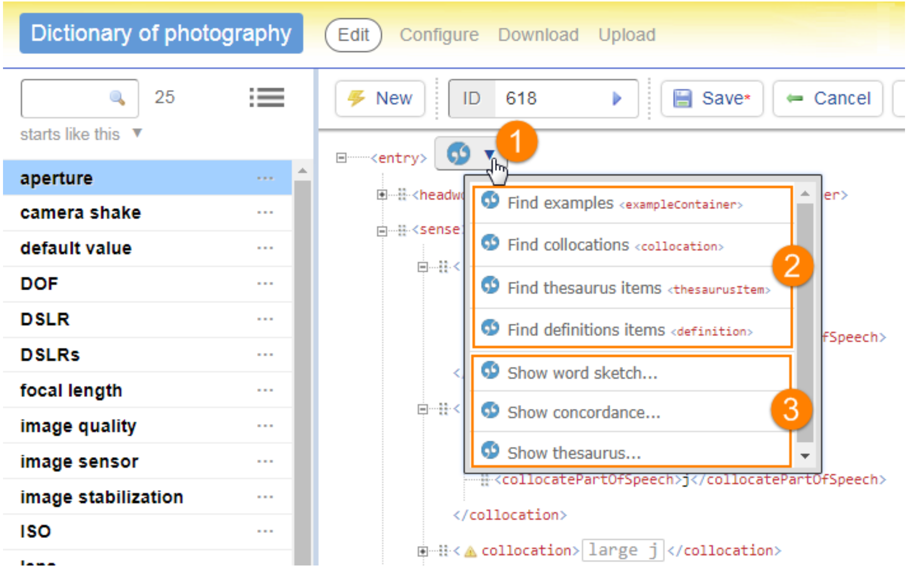
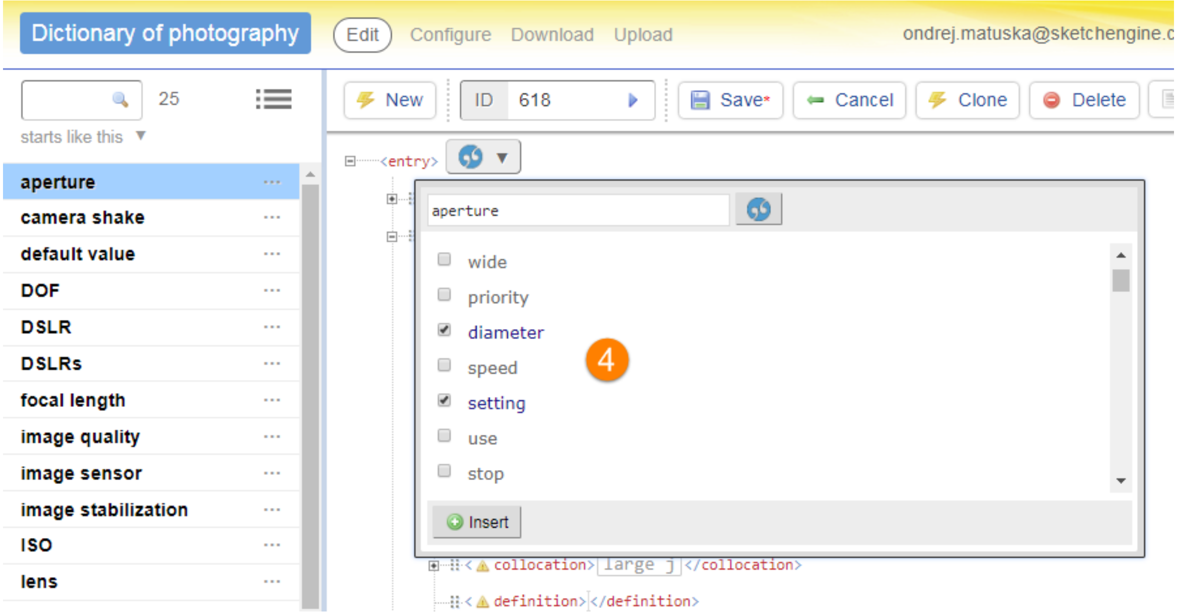


Figure 16: Interlinks between dictionary entries in Lexonomy and corresponding examples from Sketch Engine.

The dictionary editor can decide to use the Sketch Engine link **(1)** to pull data from Sketch Engine into Lexonomy **(2)** or to jump to the result screen in Sketch Engine **(3)** where all available Sketch Engine tools can be used for a more detailed analysis. The data pulled from Sketch Engine **(4)** can be revised prior to including them into the dictionary entry and they can be edited afterwards.

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.



aperture

☐ wide
☐ priority
☑ diameter
☐ speed
☑ setting
☐ use
☐ stop

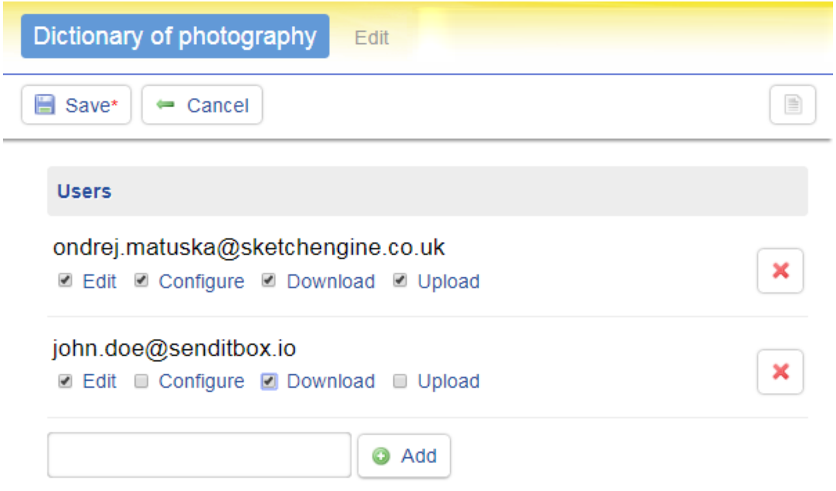
```
<collocation> large </collocation>
<definition></definition>
```

Figure 17: Lexonomy: data pulled from Sketch Engine can be edited.

3.1.2.9 Collaborative editing

Lexonomy already supports collaborative work which is vital to lexicographic projects. Users with different level of access permissions can be added to each dictionary.

There are plans to develop additional functionality to support collaborative editing and foster cooperation within the editorial team.



ondrej.matuska@sketchengine.co.uk
☑ Edit ☑ Configure ☑ Download ☑ Upload

john.doe@senditbox.io
☑ Edit ☐ Configure ☑ Download ☐ Upload

Figure 18: A list of access privileges to a dictionary in Lexonomy.

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.

3.1.2.10 Dictionary visualisation

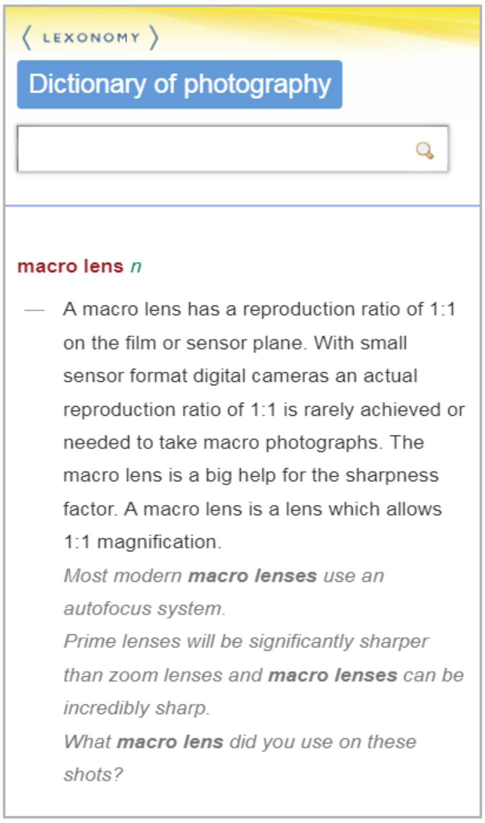
Each element within the dictionary entry has its default styling (colour, font size, the use of italics or boldface). The user can, however, override this styling to adapt it to the concrete lexicographic project. Lexonomy also supports conditional formatting and scripting which can be used to automatically adapt the visualisation of individual entries depending on the data they contain.

3.1.2.11 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 in a responsive web interface which adapt to both desktop monitors as well as the screen of mobile devices in Figure 19. The interface includes search functionality. The dictionary configuration can be used to define which dictionary entry elements should be included in the search.

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

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.



macro lens *n*

— 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. The macro lens is a big help for the sharpness factor. A macro lens is a lens which allows 1:1 magnification.

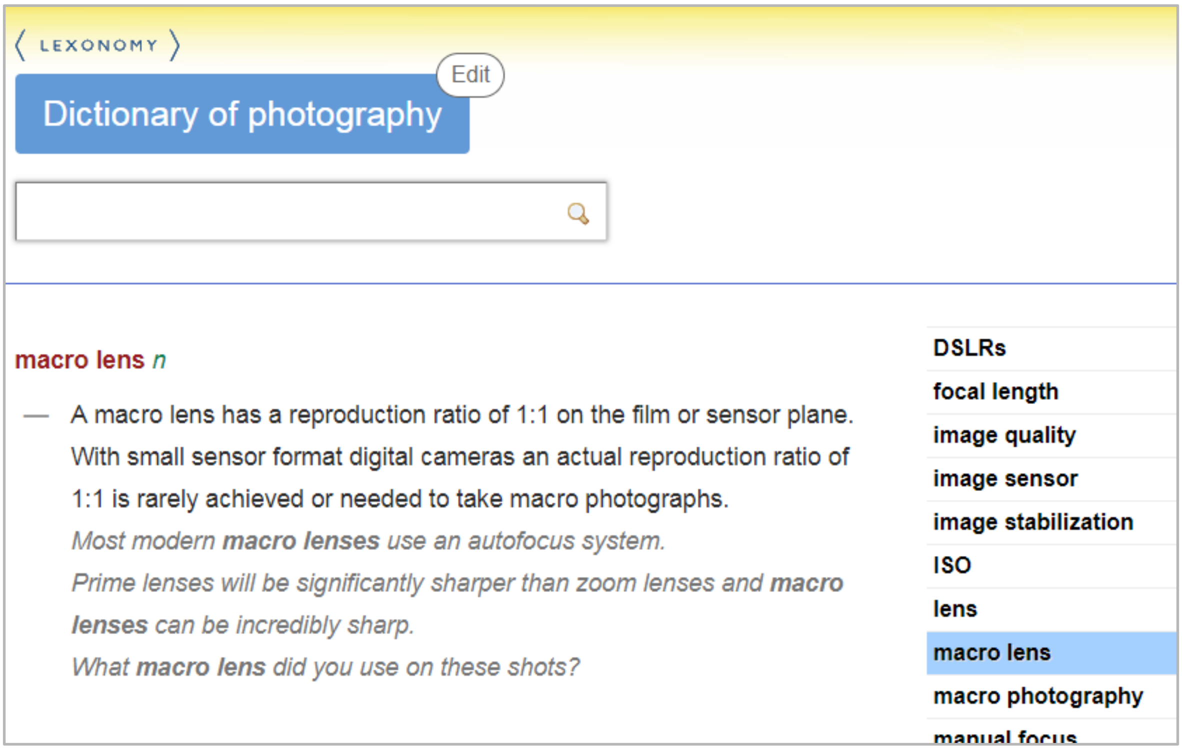
*Most modern **macro lenses** use an autofocus system.*

*Prime lenses will be significantly sharper than zoom lenses and **macro lenses** can be incredibly sharp.*

*What **macro lens** did you use on these shots?*

Figure 19: Mobile resolution of Lexonomy.

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.



macro lens *n*

— 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.

*Most modern **macro lenses** use an autofocus system.*

*Prime lenses will be significantly sharper than zoom lenses and **macro lenses** can be incredibly sharp.*

*What **macro lens** did you use on these shots?*

Figure 20: Lexonomy on desktop monitors.

3.2 Technical provisions for LEX2

Technically the whole LEX2 infrastructure is provided as a web service using secured access (HTTPS). User authentication and authorization is arranged through the eduGAIN federation network operated by the GEANT Association. As of 2019, all EU countries except for Slovakia and Bulgaria are eduGAIN partners, with these two countries being candidates for membership.[1]

The eduGAIN federation interconnects national identity federations effectively providing a Single-Sign-On (SSO) facility for researchers worldwide. In an SSO-based authentication scenario, users requesting access to service are redirected to the users' domestic institution to validate their identity as illustrated in Figure 21**.** The eduGAIN federation network manages interconnects national identity federations who manage metadata for both service providers and users. Users accessing the LEX2

[1] Detailed information for ELEXIS users is provided at https://www.sketchengine.eu/elexis/.

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.

services are first redirected to a signpost web page where they select their domestic institutions (identity provider) and then proceed with authentication.

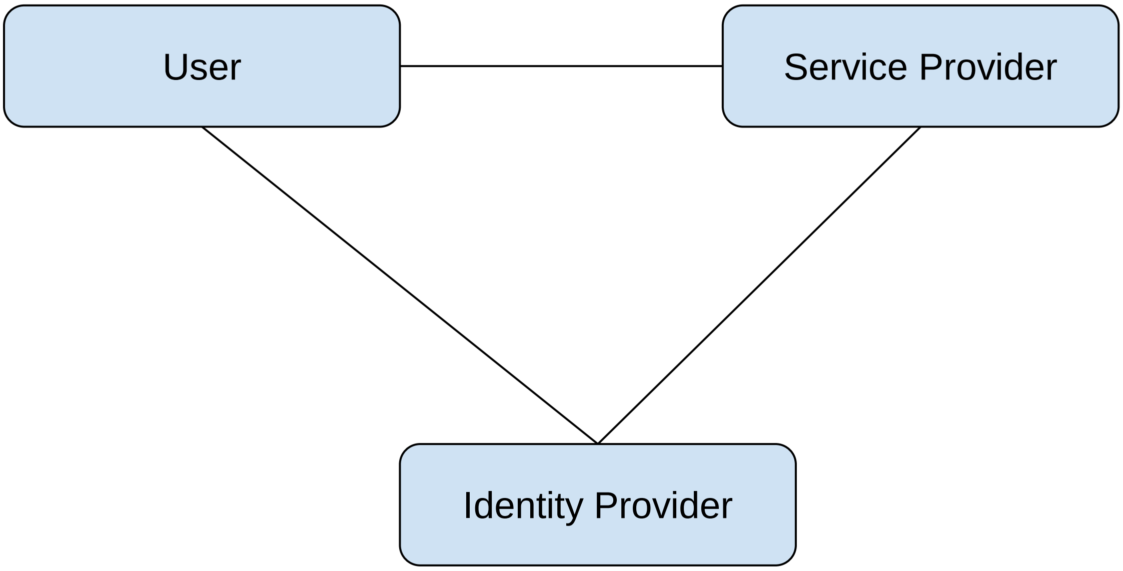


Figure 21: A simple diagram representing infrastructure of user access via SSO.

Lexical Computing is a registered service provider operating in the eduGAIN network. During the first year of the project, Sketch Engine and Lexonomy system became available as part of the service portfolio accessible throughout eduGAIN. To ease access for as many institutions as possible, Lexical Computing has acquired the ISO 27001 security certification in 2019 as well as validated its services for the GEANT Data Protection Code of Conduct, an initiative that brings the federation network framework in line with the EU General Data Protection Regulation (GDPR).

The LEX2 infrastructure is operated solely by Lexical Computing and hosted in a dedicated cluster of servers in a private data centre. During 2019 this infrastructure has been upgraded to allow for a distributed storage and therefore parallel search and management of the corpora hosted in Sketch Engine. Changes in the implementation of the Sketch Engine associated with this innovation are described in [4].

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.

3.3 Statistics of usage of LEX2

The access to the infrastructure was launched on 1 April 2018. The information campaign started as early as January 2018 and by the end of April 2018, 129 academic institutions (mainly universities) had been granted access to Sketch Engine.

3.3.1 Number of institutions

A total of 337 institutions gained access until the end of 2019. Figure 22 gives an overview of the number of institutions enjoying the access to the ELEXIS infrastructure.

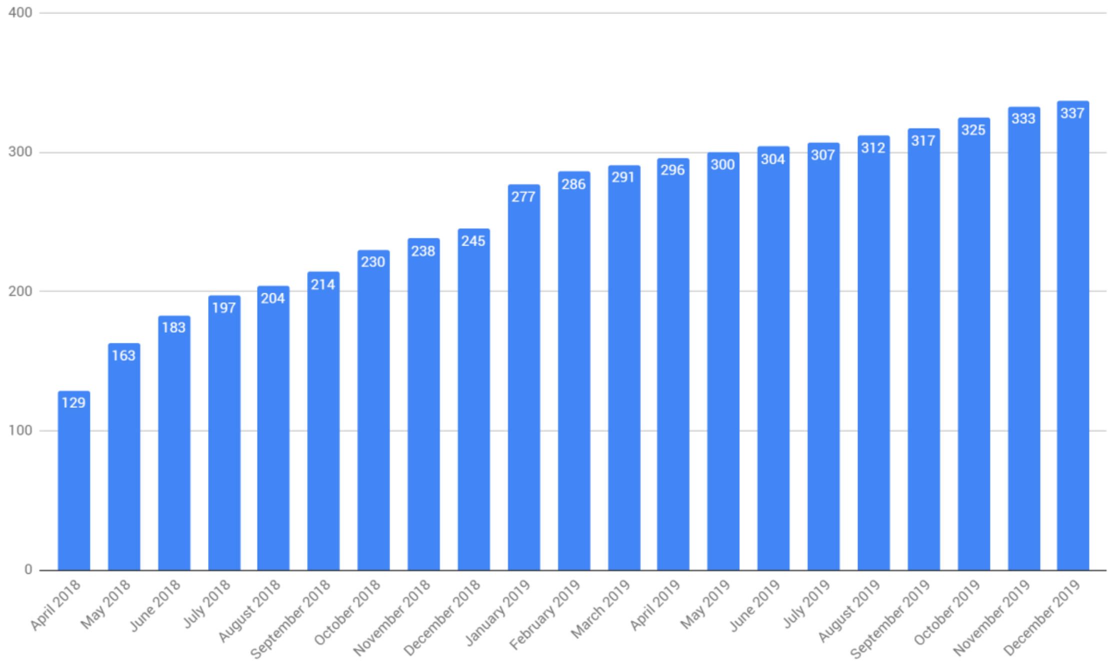


Figure 22: A total number of institutions joined the ELEXIS project per each month in 2018.

The growth follows a steady path with several applications received every month even after two years of the project. There are many institutions aware of the free access to the infrastructure and they are in the process of going through their internal procedures of submitting their enrolment forms. More institutions are still expected to join the scheme.

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.

3.3.2 Country representation

Institutions from a total of 26 EU countries had shown interest in the ELEXIS infrastructure until the end of 2019 and submitted their enrolment forms (Figure 23).

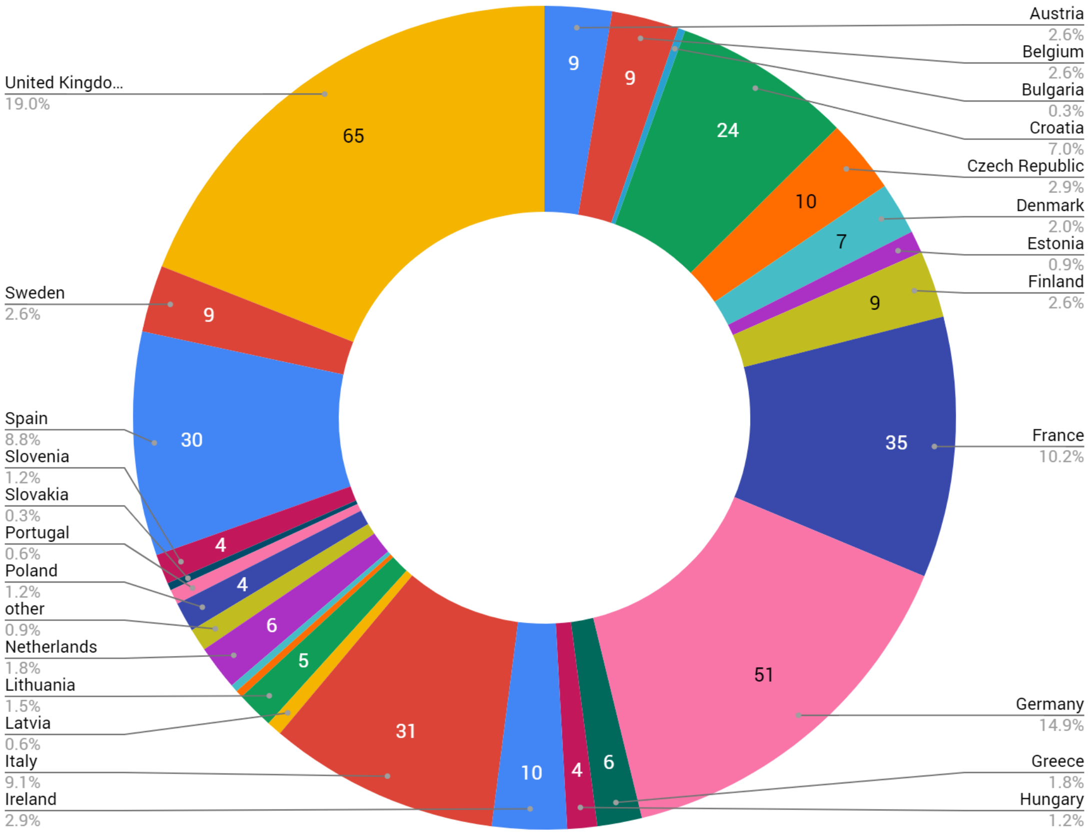


Figure 23: A number of institutions joined the ELEXIS project per country.

The only two EU countries not represented are Romania and Cyprus. Romania is a country without a functioning national identity federation enabling their institution to gain access using the standard setup procedure. Lexical Computing offers interested Romanian institutions alternative methods of gaining access. The same methods were used to connect institutions in Slovakia and Bulgaria where the identity federation is also non-existent. Academic institutions in Cyprus have not shown interest so far.

In addition, three institutions from outside of the EU were granted access as part of their observer status within ELEXIS.

3.3.3 Number of user accounts

The number of user accounts saw a notable rise in 2019 and reached 250 % as seen in Figure 24. In the light of the increase in the number of institutions reaching 30 %, it is fair to assume a favourable trend of the institutions making a more intensive use of the infrastructure and promoting the access within their institutions.

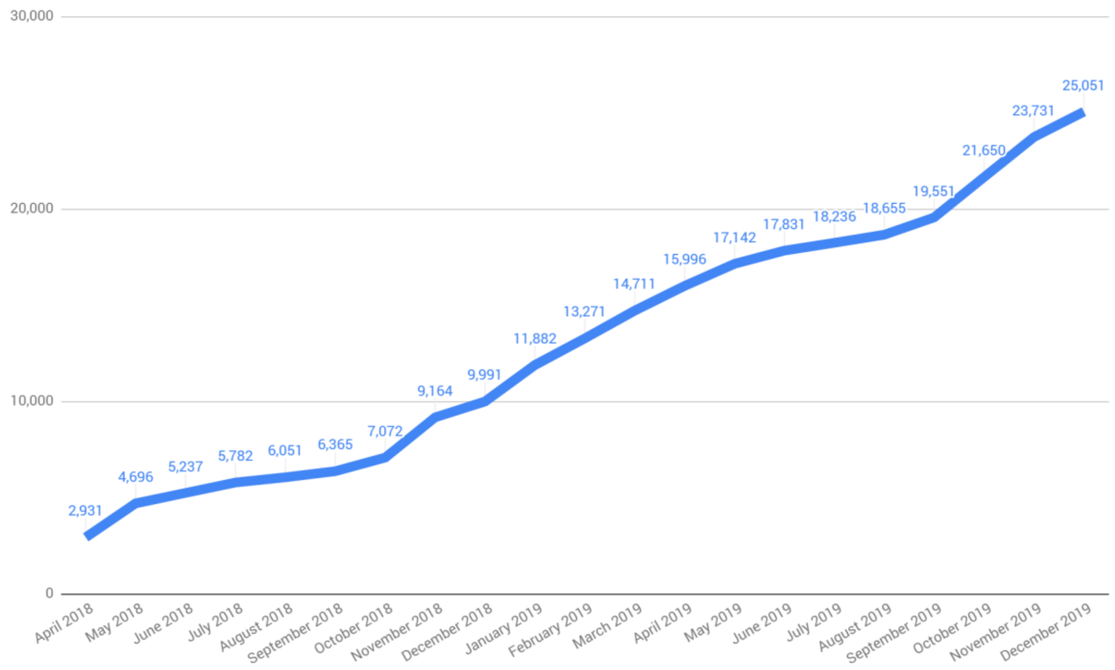


Figure 24: A total number of new users who gained access to Sketch Engine via the ELEXIS project.

3.3.4 User-hours

The intensity of use is measured in user-hours. A user-hour is defined as an hour (a 60-minute period) in which the user made at least one request. A request is defined as a mouse click which generates some activity on the screen such as clicking on the SEARCH button, changing the view settings, applying a filter etc.

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.

In the period from 1 January 2019 and 31 December 2019, users taking advantage of the ELEXIS-funded access to the infrastructure generated a total of **450,039 user-hours**, see Figure 25.

The drops in the intensity of use correspond to the predictable events such as holidays. The peaks correspond to the end of term when students finish their assignments, and when the winter term is in full swing.

The statistics clearly demonstrates that institutions make a more intensive use of Sketch Engine since the number of user-hours increased proportionally more than the number of users. This leads us to believe that not only are there more users but they also spend more time working with the infrastructure.

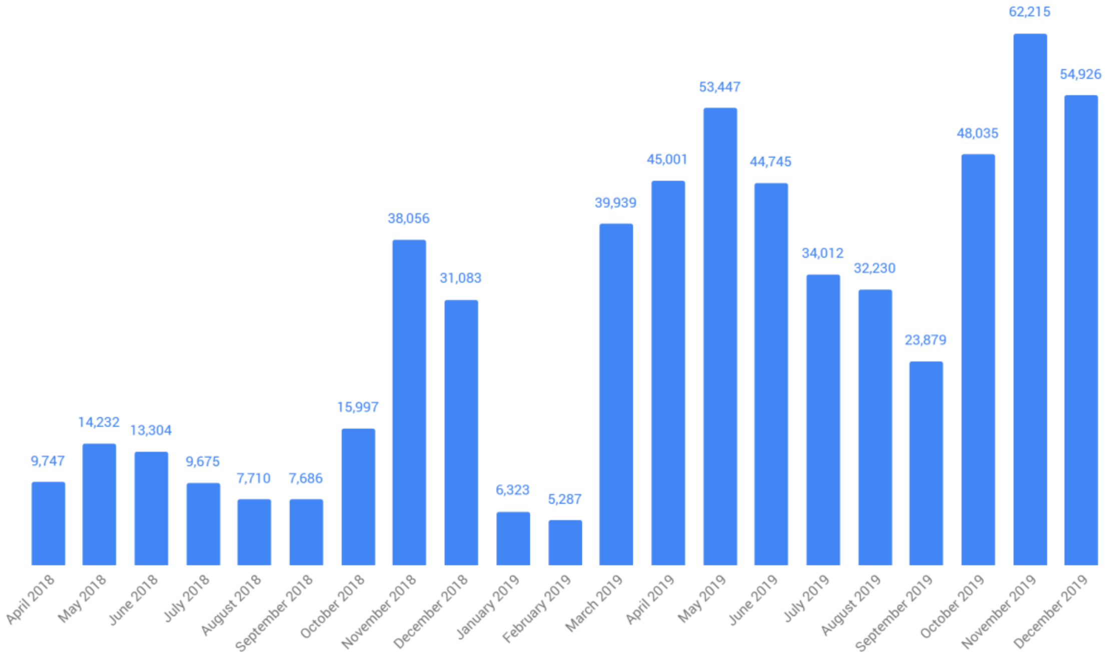


Figure 25: User-hours in each month since the project start.

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 Part 2: Platform for access to retrodigitised resources (LEX3)

The LEX3 infrastructure will consist of Dictionary Viewer Module which is developed by Trier Centre for Digital Humanities. As of M24 the moduse is still in development and is expected to be included in the infrastructure by M36.

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 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] KILGARRIFF, Adam, Miloš HUSÁK, Katy MCADAM, Michael RUNDELL and Pavel RYCHLÝ. GDEX: Automatically finding good dictionary examples in a corpus. In BERNAL Elisenda and Janet DeCESARIS *Proceedings of the 13th EURALEX International Congress*. Barcelona: Pompeu Fabra University, 2008, p. 425–432.

[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] RÁBARA, Radoslav, Pavel RYCHLÝ, Ondřej HERMAN and Miloš JAKUBÍČEK. Accelerating Corpus Search Using Multiple Cores. In BAŃSKI Piotr, Marc KUPIETZ, Harald LÜNGEN, Paul RAYSON, Hanno BIBER, Evelyn BREITENEDER, Simon CLEMATIDE, John MARIANI, Mark STEVENSON, Theresa SICK. *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section*. Mannheim: Institut für Deutsche Sprache, 2017. p. 30–34.