

D8.4

Periodic assessment
of LEX1, LEX2 and
LEX3

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


Date: 31. 7. 2022

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D8.1 Periodic assessment of LEX1, LEX2 and LEX3 –
first report



Deliverable Number: D8.4
Dissemination Level: Public
Delivery Date: 31. 7. 2022
Version: 1.0
Author(s): Miloš Jakubíček,
Ondřej Matuška,
Michal Cukr, 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
---------	------	------------------	-----------------------

Table of Contents

1	Part 1: Lexicographic data integration platform (LEX1).....	1
2	Part 2: Platform for dictionary drafting from corpora (LEX2).....	3
2.1	Tools and services in LEX2.....	3
2.1.1	SketchEngine.....	3
2.1.2	Lexonomy.....	13
2.1.2.1	Push model.....	15
2.1.2.2	Dictionary templates.....	15
2.1.2.3	Editing the dictionary.....	16
2.1.2.4	Dual editing interface.....	16
2.1.2.5	User manual.....	17
2.1.2.6	Flags.....	17
2.1.2.7	Unicode characters.....	17
2.1.2.8	Pull model.....	17
2.1.2.9	Collaborative editing.....	19
2.1.2.10	Dictionary visualisation.....	20
2.1.2.11	Publishing the dictionary.....	20
2.2	Technical provisions for LEX2.....	21
2.3	Statistics of usage of LEX2.....	22
2.3.1	Number of institutions.....	23
2.3.2	Country representation.....	23
2.3.3	Number of user accounts.....	24
2.3.4	User-hours.....	25
3	Part 3: Platform for access to retrodigitised resources (LEX3).....	27
4	References.....	28

List of Figures

1	Part 1: Lexicographic data integration platform (LEX1)	1
2	Part 2: Platform for dictionary drafting from corpora (LEX2)	3
2.1	Tools and services in LEX2	3
2.1.1	SketchEngine	3
2.1.2	Lexonomy	13
2.1.2.1	Push model	15
2.1.2.2	Dictionary templates	15
2.1.2.3	Editing the dictionary	16
2.1.2.4	Dual editing interface	16
2.1.2.5	User manual	17
2.1.2.6	Flags	17
2.1.2.7	Unicode characters	17
2.1.2.8	Pull model	17
2.1.2.9	Collaborative editing	19
2.1.2.10	Dictionary visualisation	20
2.1.2.11	Publishing the dictionary	20
2.2	Technical provisions for LEX2	21
2.3	Statistics of usage of LEX2	22
2.3.1	Number of institutions	23
2.3.2	Country representation	23
2.3.3	Number of user accounts	24
2.3.4	User-hours	25
3	Part 3: Platform for access to retrodigitised resources (LEX3)	27
4	References	28

1 Part 1: Lexicographic data integration platform (LEX1)

This part of the infrastructure is covered with the task T8.1 which started in M12. The LEX1 infrastructure is hosted and operated by Jozef Stefan Institute which has led the negotiations with other project partners on suitable tools for securing all parts of this infrastructure. As of M36 the infrastructure consists of a server providing cloud services for hosting of all project data (corpora, dictionaries and tools part of LEX1), **Elexifier** conversion module and **Lexonomy** editing and publishing module. Linking and enriching modules are in development and expected to be launched in M42.

The central LEX1/LEX2/(LEX3) infrastructure was launched in March 2020 and consists of six interlinked service modules which draw on tools already developed within LEX2.

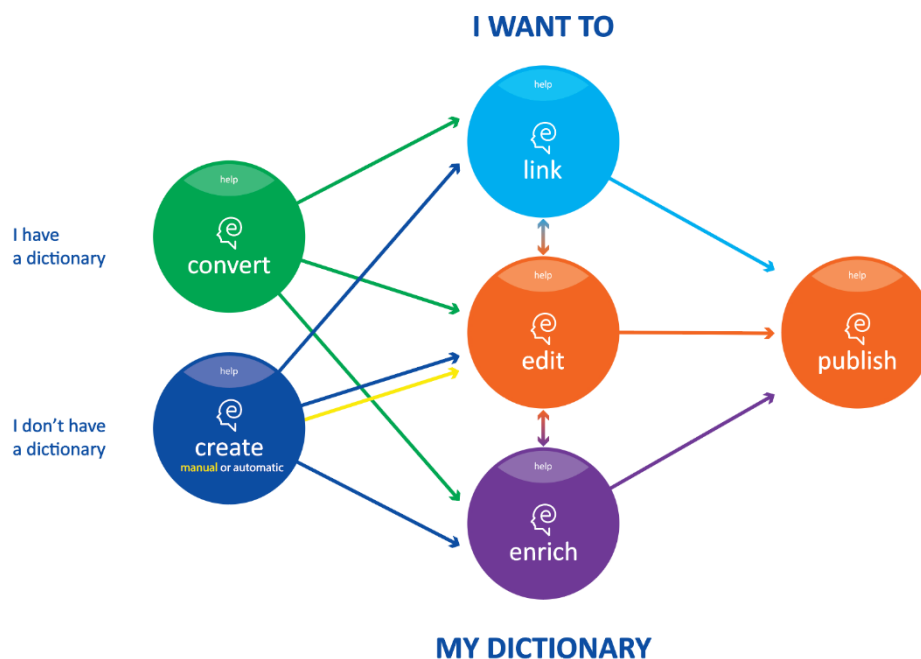


Figure 1: LEX 1 modules.

The modules, shown in Figure 1, are:

D8.3 Periodic assessment of LEX1, LEX2 and LEX3

- 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]. The tool design is described in D1.3 Tools for the automatic segmentation and identification of lexicographic content. The use of the tool is described in D6.3 Intermediate Interoperability Report.
- 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]. See also Part 2 in this document.
- EDIT module provides the option of uploading and compilation of the dictionary in the Lexonomy dictionary writing system. [ALREADY AVAILABLE AS PART OF LEX1] See Part 2.
- 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 M42]
- LINK module provides services of linking dictionary entries or parts of entries with other monolingual and multilingual lexicographic resources. [WILL BE AVAILABLE IN M42]
- 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. [LEXONOMY IS AVAILABLE; DICTIONARY PORTAL WILL BE AVAILABLE IN M48]

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



2 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

2.1 Tools and services in LEX2

2.1.1 SketchEngine



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



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

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.

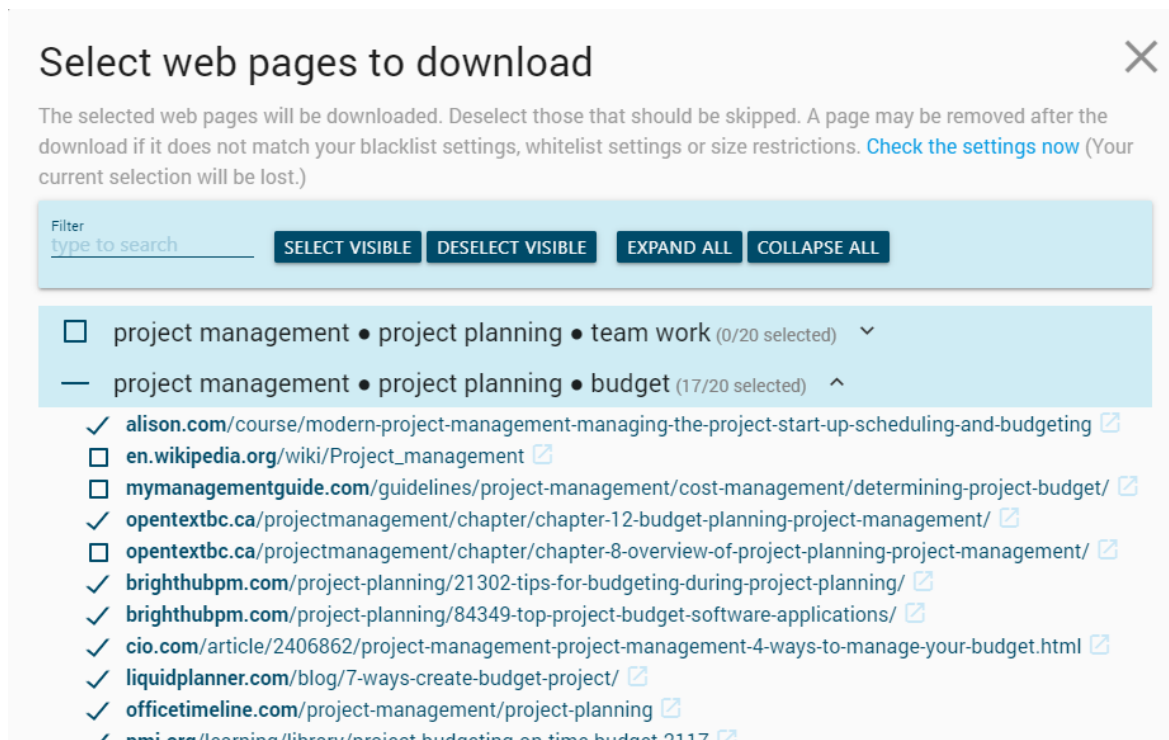


Figure 2: An interface allowing the user to approve the web resources identified by Sketch Engine.

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.

D8.3 Periodic assessment of LEX1, LEX2 and LEX3



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*.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

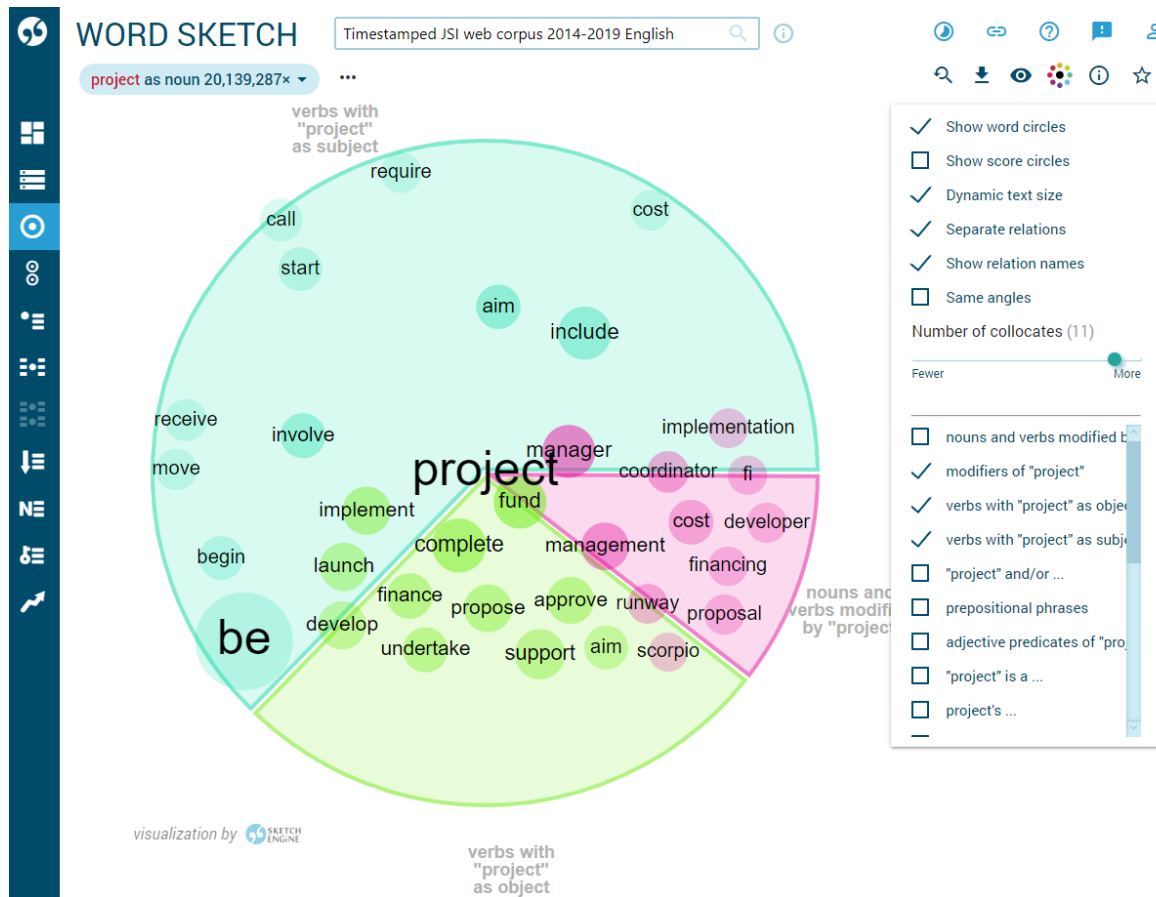


Figure 4: Results for the noun “project” from the Word sketch feature of Sketch Engine.

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]



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

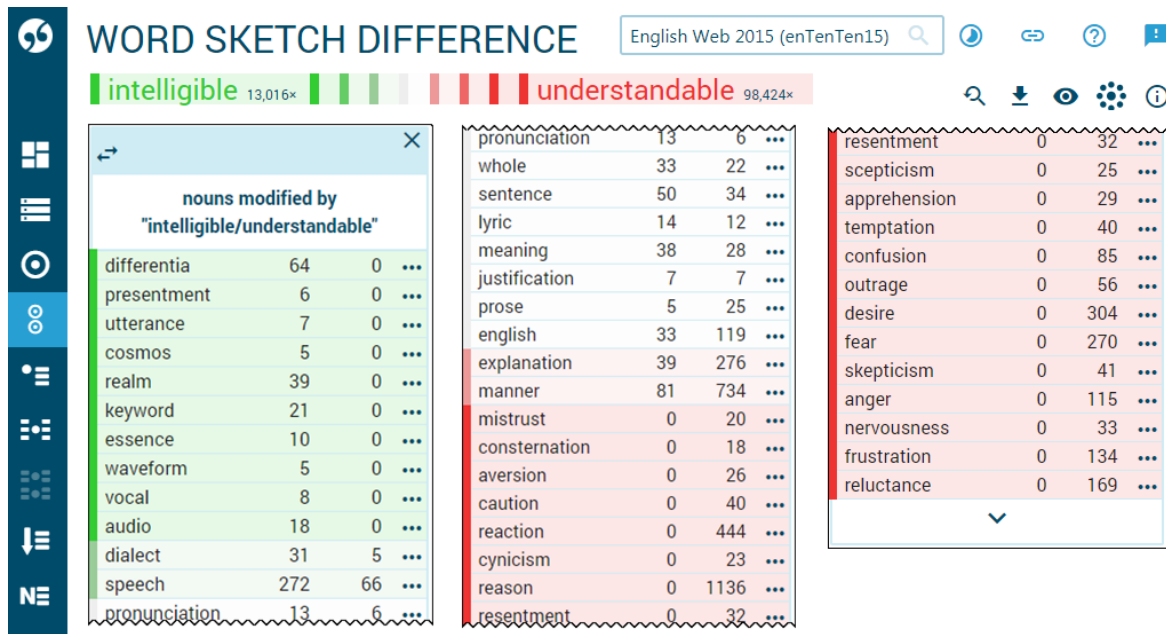


Figure 5: A layout of the word sketch difference tool for the adverbs "intelligible" and "understandable".

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

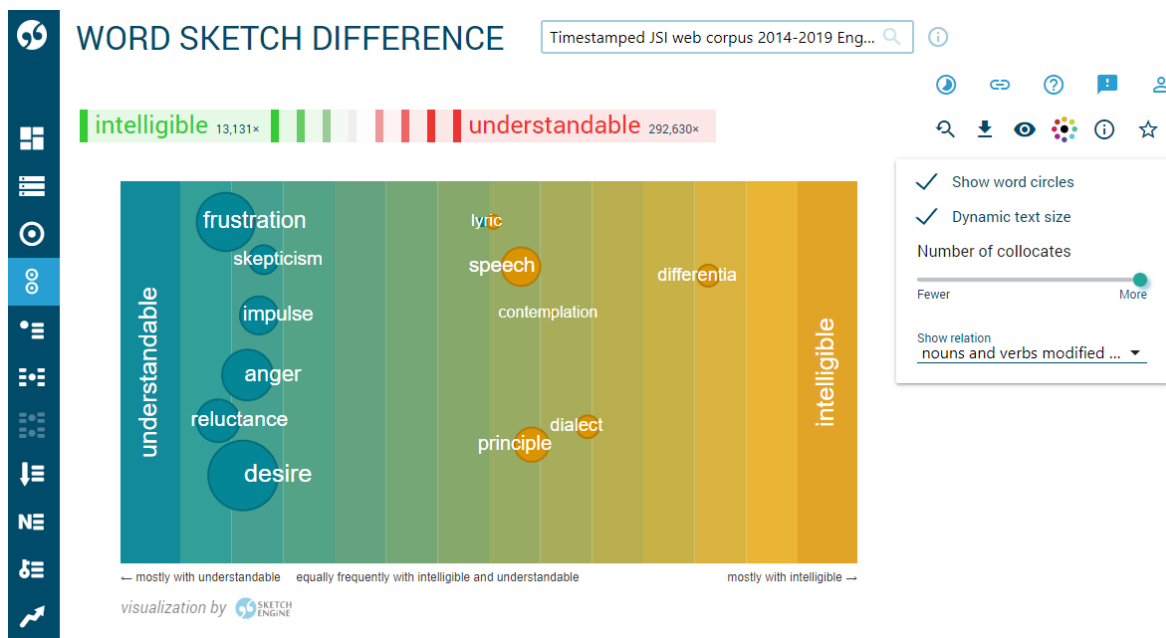


Figure 6: Visualisation of the differences in collocations modifying the words "intelligible" and "understandable".



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

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]

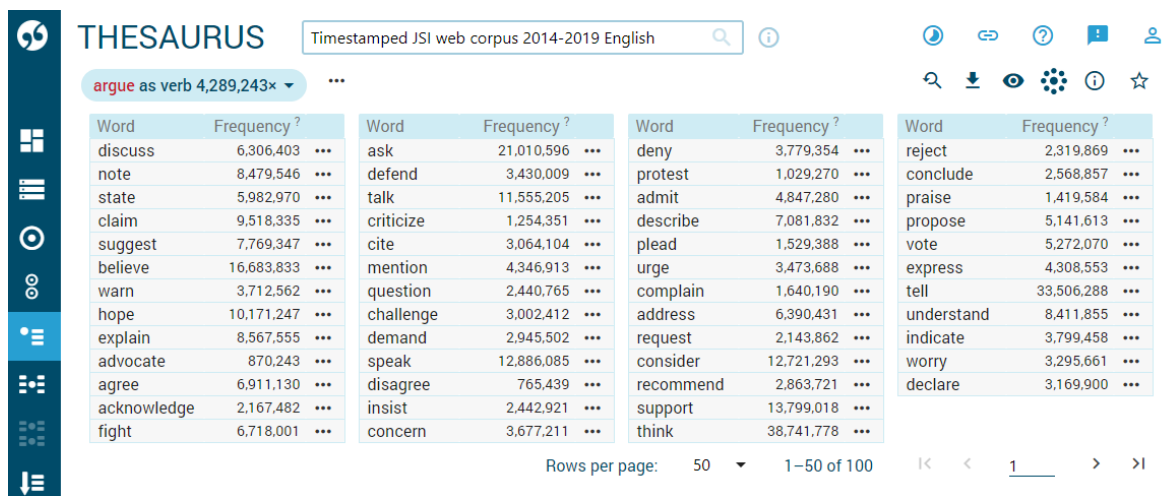


Figure 7: A thesaurus list for the word "argue" with words sorted by the similarity to argue based on the principles of distributional semantics.

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.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

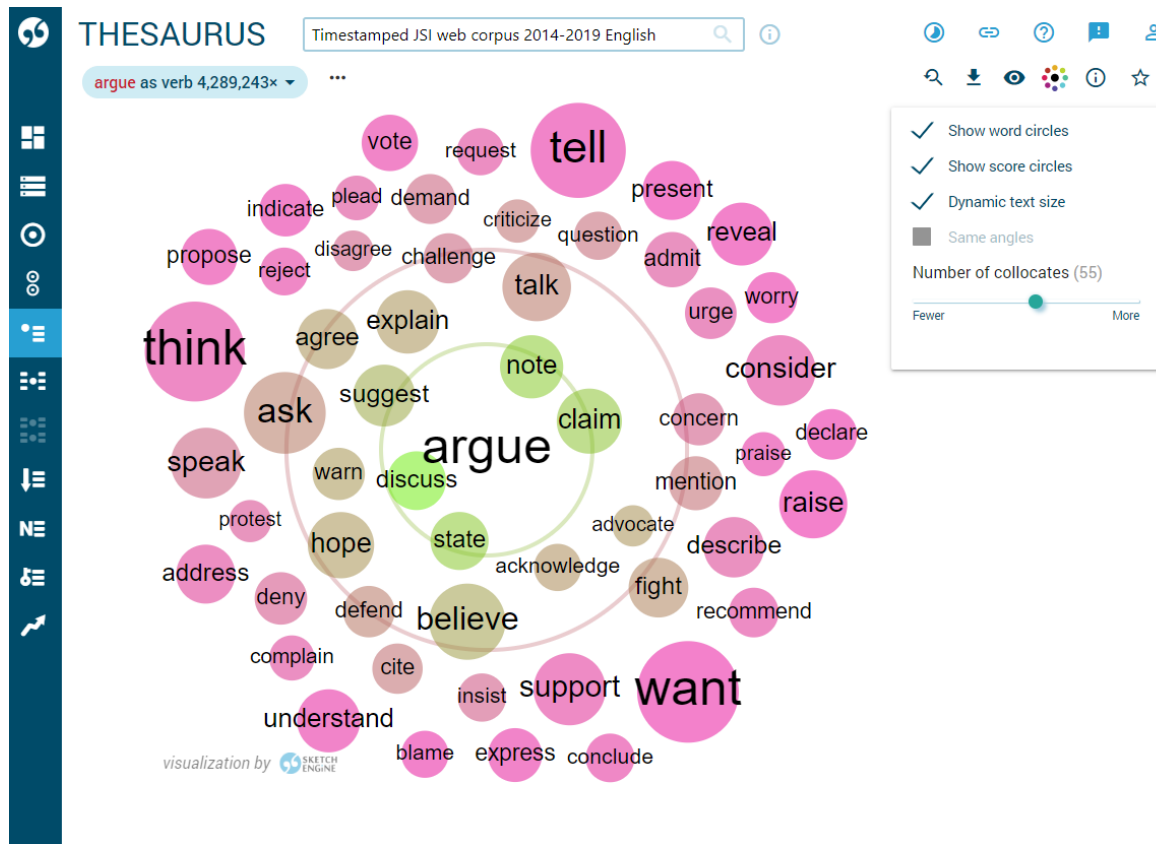
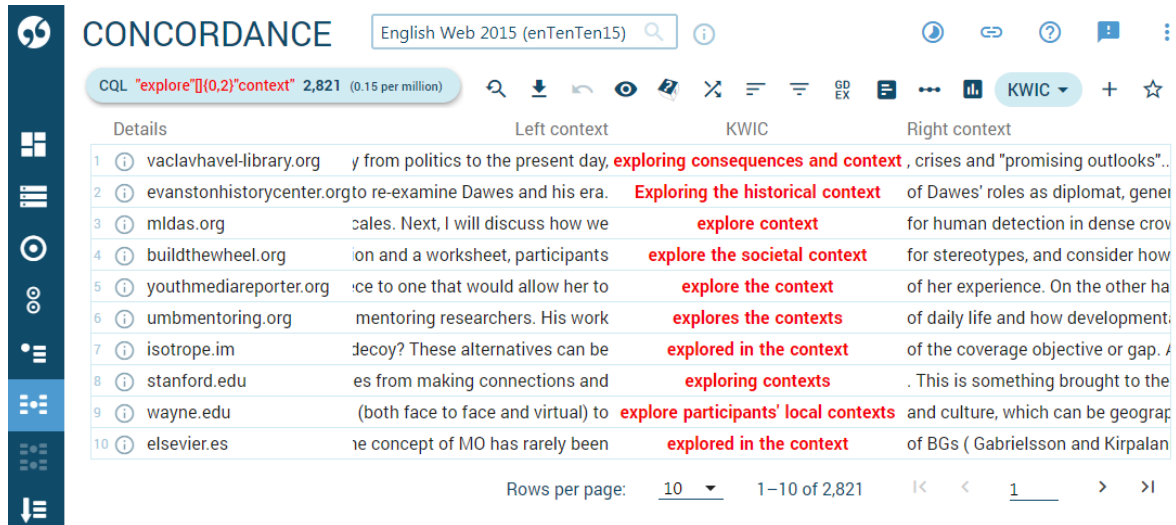


Figure 8: A visualisation of the thesaurus for the word "argue".

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 9) 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.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3



The screenshot shows the CONCORDANCE interface with the following details:

- Search engine: English Web 2015 (enTenTen15)
- Query: CQL "explore" (0,2)"context" 2,821 (0.15 per million)
- Table columns: Details, Left context, KWIC, Right context
- Table rows (1-10):

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, gene
3 midas.org	cales. Next, I will discuss how we	explore context	for human detection in dense crow
4 buildthewheel.org	ion and a worksheet, participants	explore the societal context	for stereotypes, and consider how
5 youthmediareporter.org	nce 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	Jecoy? These alternatives can be	explored in the context	of the coverage objective or gap. A
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	re concept of MO has rarely been	explored in the context	of BGs (Gabrielsson and Kirpalan
- Rows per page: 10 (dropdown)
- Page 1 of 1 (1-10 of 2,821)

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.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

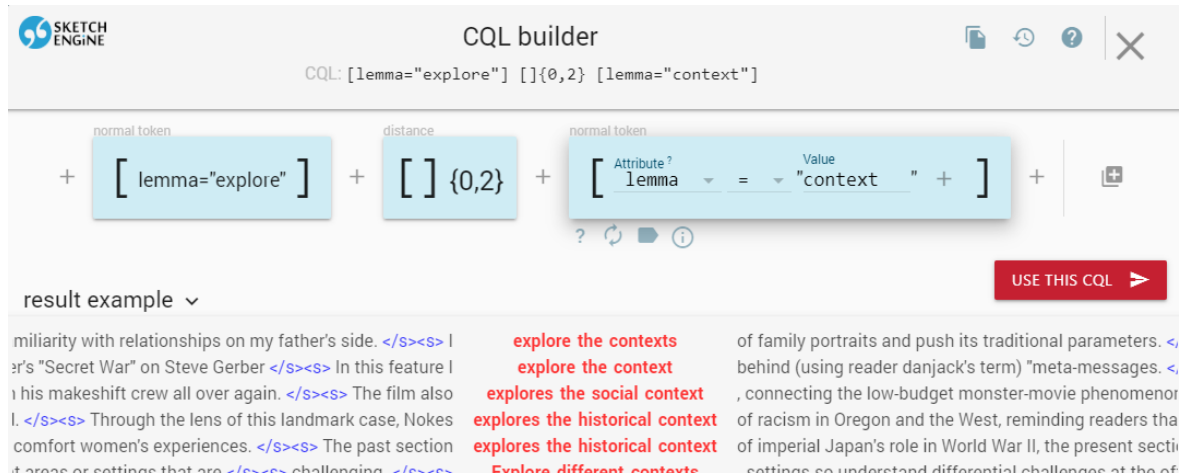


Figure 10: The CQL builders with the query used to generate the concordance in Figure 8.

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



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

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.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

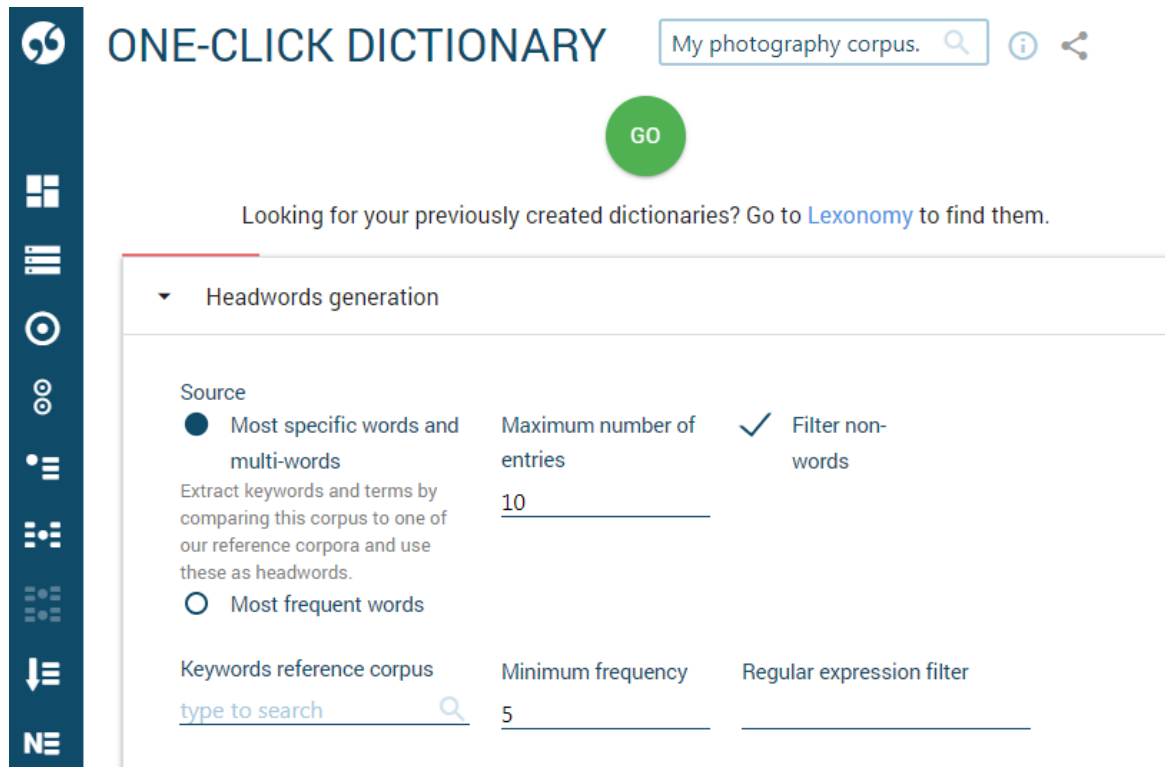


Figure 11: OneClick Dictionary – setting up the building of a new dictionary draft from a corpus.

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.

2.1.2 Lexonomy



access on www.lexonomy.eu



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

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.

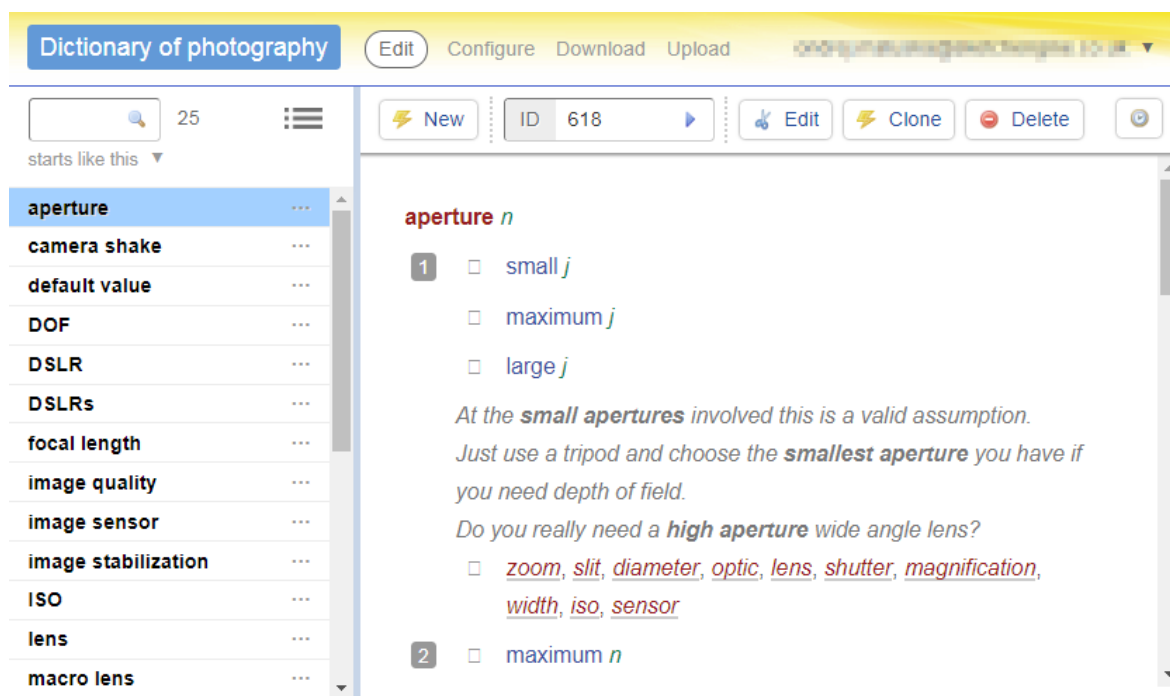


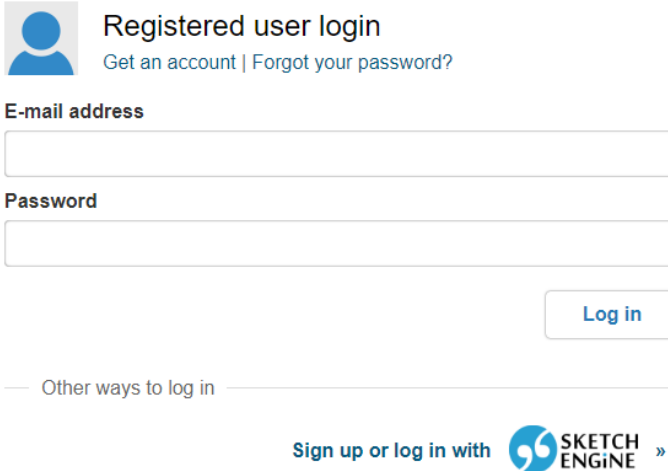
Figure 12: A dictionary entry in Lexonomy.

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).



D8.3 Periodic assessment of LEX1, LEX2 and LEX3



Registered user login
Get an account | Forgot your password?

E-mail address

Password

[Log in](#)

Other ways to log in


Sign up or log in with  SKETCH ENGINE »

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

2.1.2.1 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.

2.1.2.2 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 restrictions can be defined by the user. This ensures consistency across all dictionary entries. Each

15



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

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.

2.1.2.3 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.

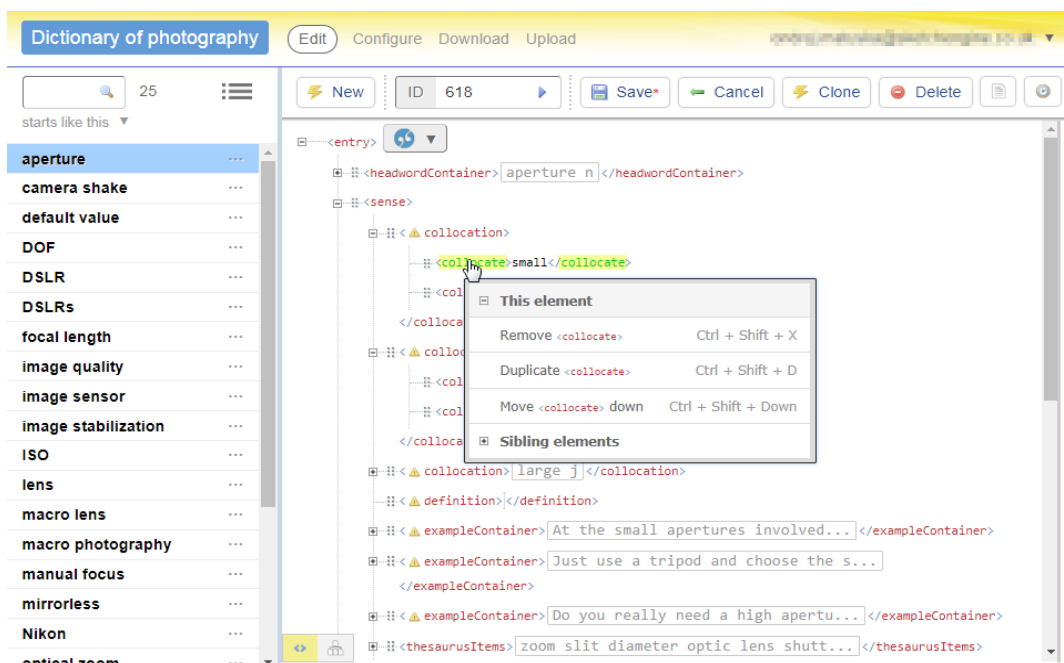


Figure 14: Editing particular attributes of a dictionary entry within Lexonomy.

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

2.1.2.4 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 15)

D8.3 Periodic assessment of LEX1, LEX2 and LEX3

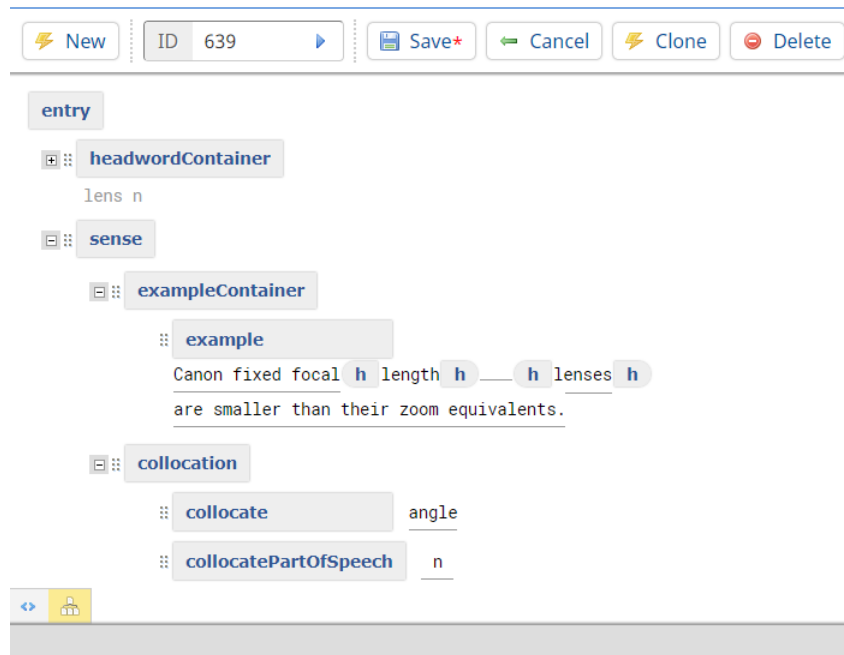


Figure 15: Dictionary entry editor - visual layout.

2.1.2.5 User manual

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

2.1.2.6 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.

2.1.2.7 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)

2.1.2.8 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-



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

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

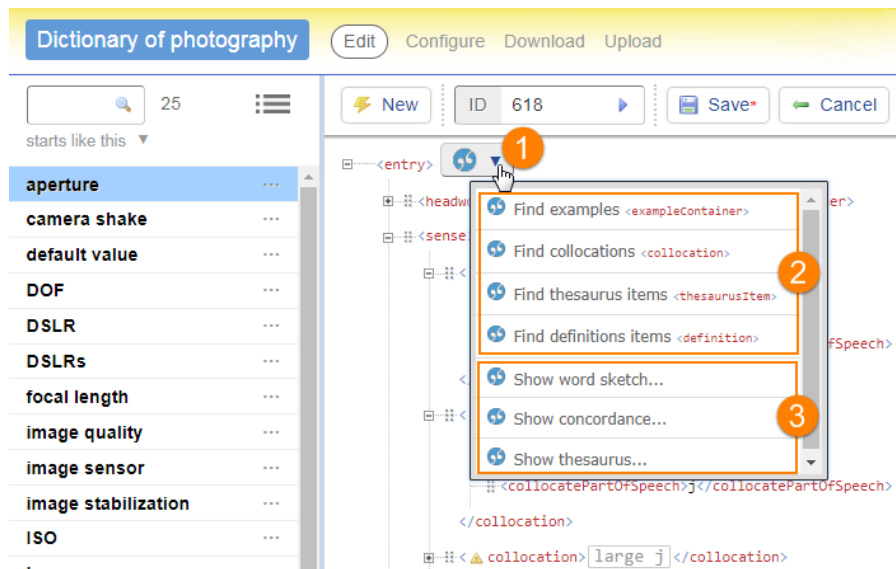


Figure 16: Interlinks between dictionary entries in Lexonomy and corresponding examples from Sketch Engine.

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.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

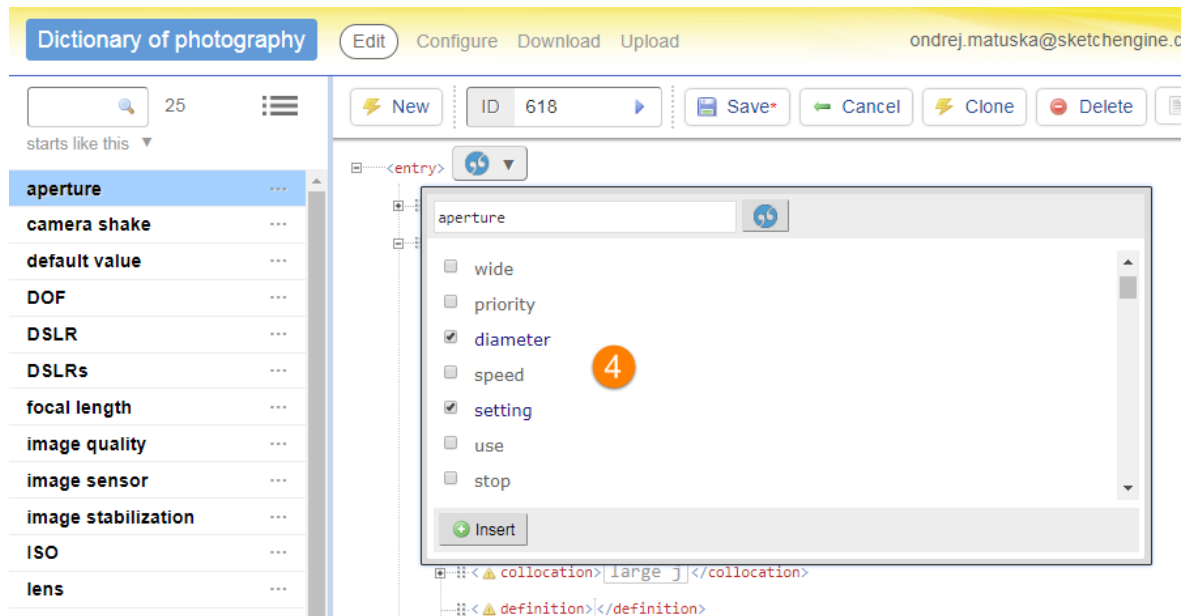


Figure 17: Lexonomy: data pulled from Sketch Engine can be edited.

2.1.2.9 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.

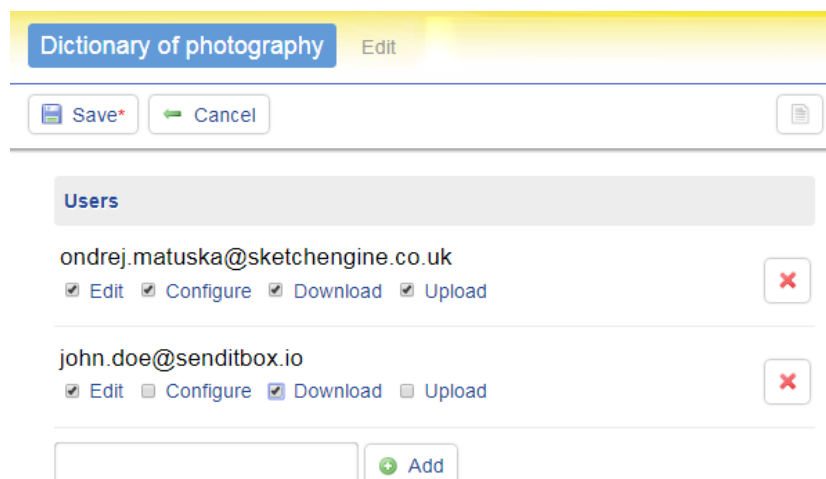


Figure 18: A list of access privileges to a dictionary in Lexonomy.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

2.1.2.10 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.

2.1.2.11 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.



The screenshot shows a web interface for 'LEXONOMY'. At the top, there is a navigation bar with the text '< LEXONOMY >'. Below this is a blue button labeled 'Dictionary of photography'. Underneath the button is a search bar with a magnifying glass icon. The main content area displays a dictionary entry for 'macro lens' with a red 'n' indicating a noun. The entry includes a definition: '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.' Below the definition are three italicized sentences: 'Most modern **macro lenses** use an autofocus system.', 'Prime lenses will be significantly sharper than zoom lenses and **macro lenses** can be incredibly sharp.', and 'What **macro lens** did you use on these shots?'

D8.3 Periodic assessment of LEX1, LEX2 and LEX3

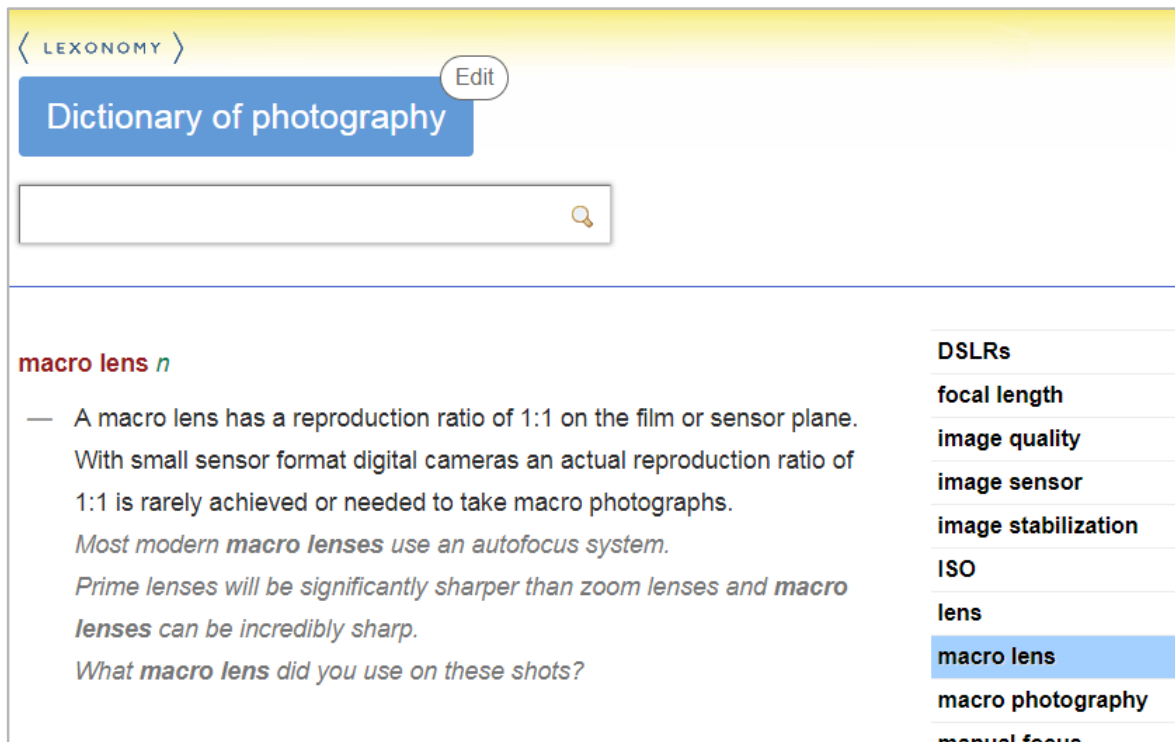


Figure 20: Lexonomy on desktop monitors.

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

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



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

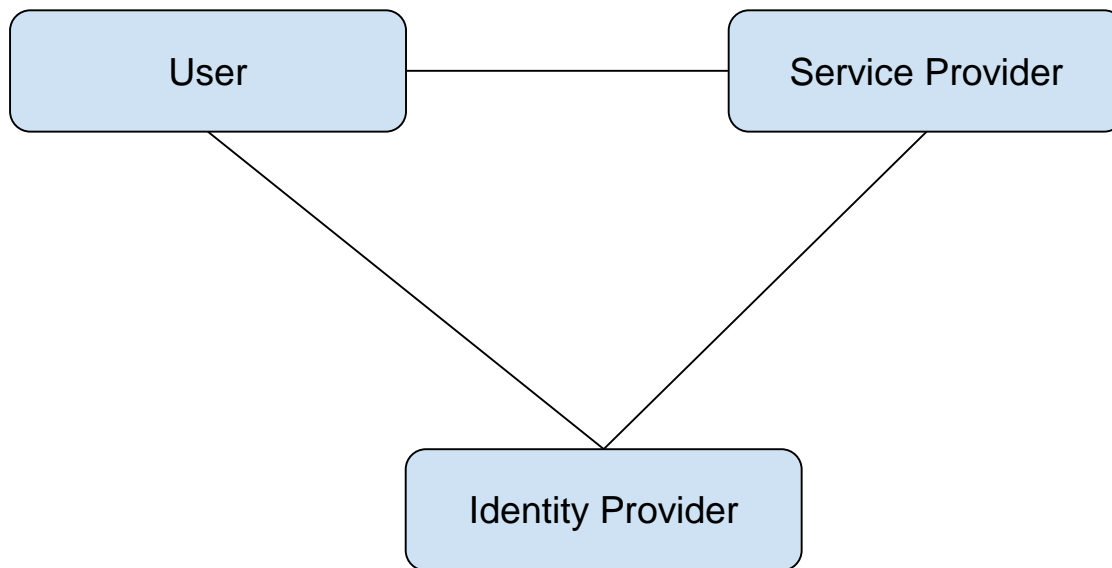


Figure 21: A simple diagram representing infrastructure of user access via SSO.

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 was 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].

2.3 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.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

2.3.1 Number of institutions

A total of 461 institutions gained access until the end of March 2022. Figure 22 gives an overview of the number of institutions enjoying the access to the ELEXIS infrastructure.

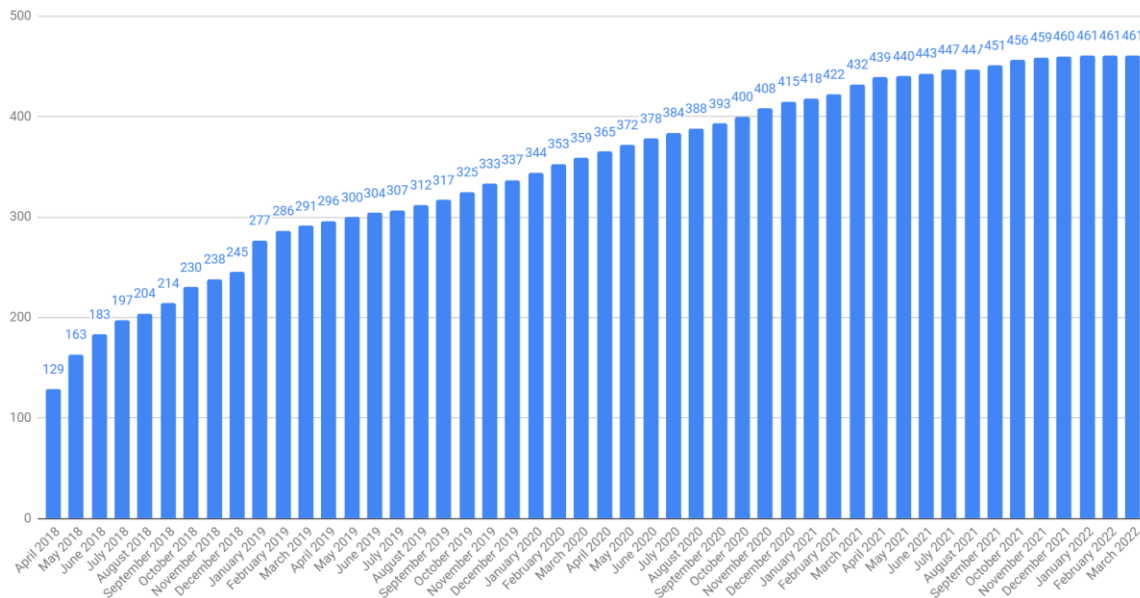


Figure 22: A total number of institutions joined the ELEXIS project per month.

The growth follows a steady path with a flat curve towards the end when institutions no longer considered it to be worthwhile going through the setup procedure with only a couple of months ahead.

2.3.2 Country representation

Institutions from all 27 EU countries had shown interest in the ELEXIS infrastructure until the end of March 2022, submitted their enrolment forms and their access had been set up (Figure 23). In addition, ELEXIS observers from 6 non-EU countries (Albania, Canada, Norway, Serbia, United Kingdom and United States) had also been granted access.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

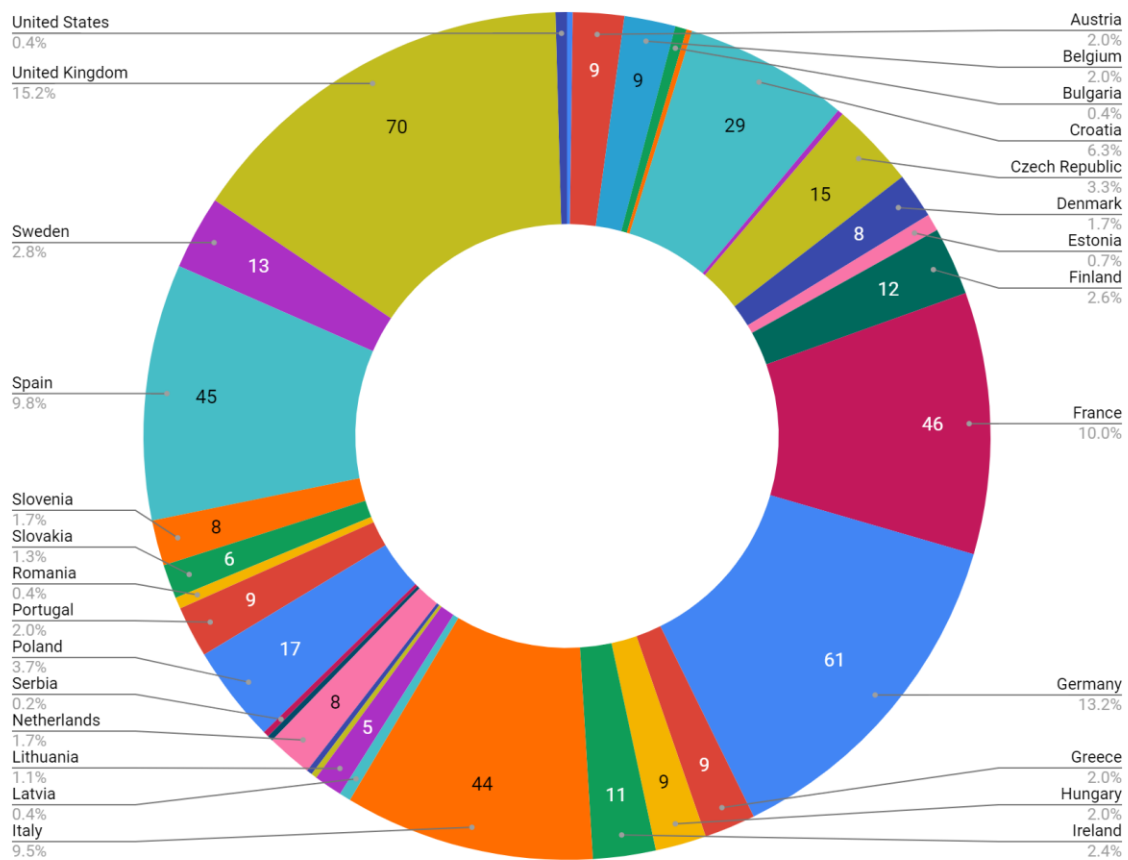


Figure 23: A number of institutions joined the ELEXIS project per country.

The United Kingdom kept the access until the end under the conditions of the Brexit withdrawal agreement. Access from outside the EU was only possible for institutions which fulfilled the conditions for acquiring the ELEXIS observer status.

2.3.3 Number of user accounts

The number of user accounts reached a total of 71,929 of students, teachers and researchers who registered their Sketch Engine accounts in the period between April 2018 and March 2022. The increase was strongest in the last year of the project when universities put Sketch Engine to the most intensive use. This is best demonstrated by the fact that in the last 15 months (January 2021 – March 2022), as many as 28,631 users started using the infrastructure. This is 40% of the total number of users who used the infrastructure in the 48-month duration of the access, see Figure 24.



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

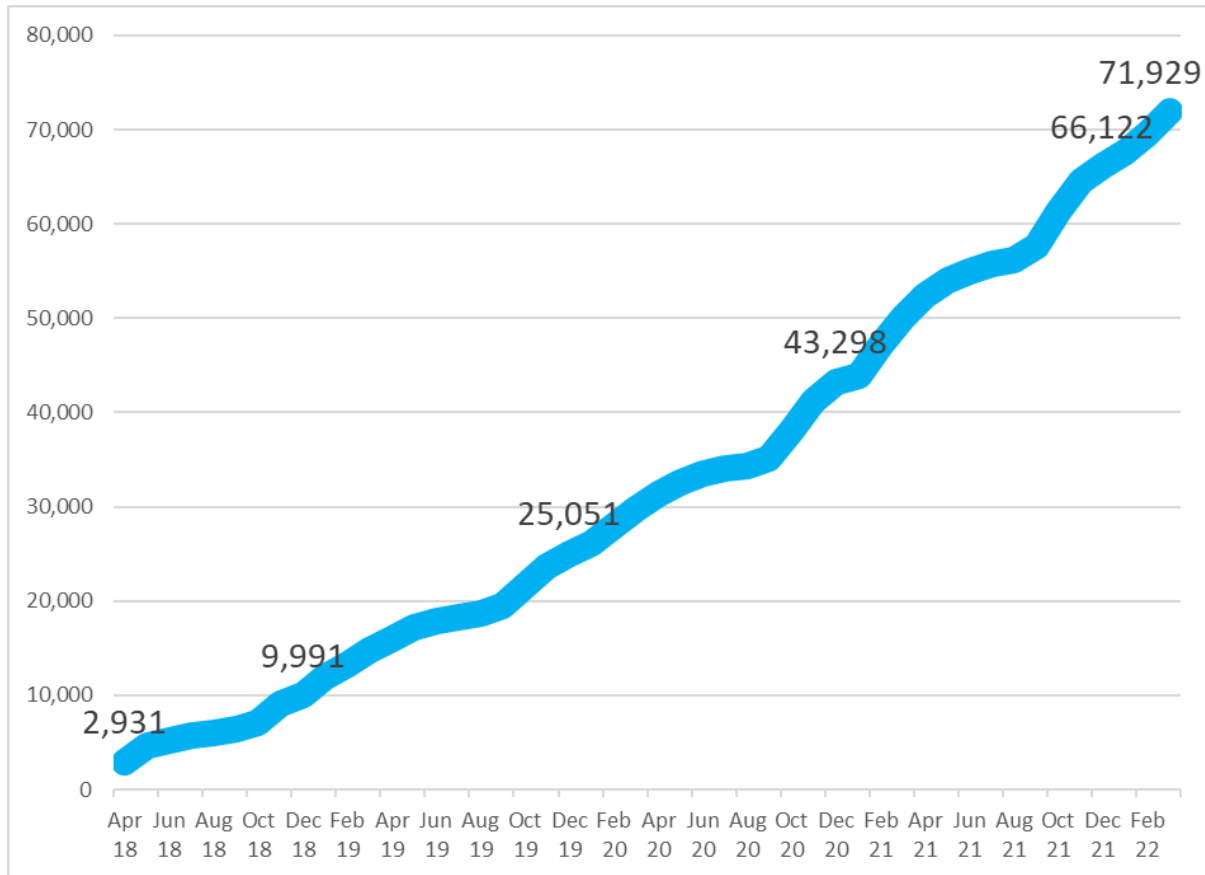


Figure 24: A total number of new users who gained access to Sketch Engine via the ELEXIS project.

2.3.4 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.

In the course of the access provided by ELEXIS, users generated 3,460,207 hours of use, 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 many times more than the number of users. This



D8.3 Periodic assessment of LEX1, LEX2 and LEX3

leads us to believe that not only are there more users but they also spend more time working with the infrastructure.

The COVID19 measures and lockdowns did not affect the intensity of use of the infrastructure.

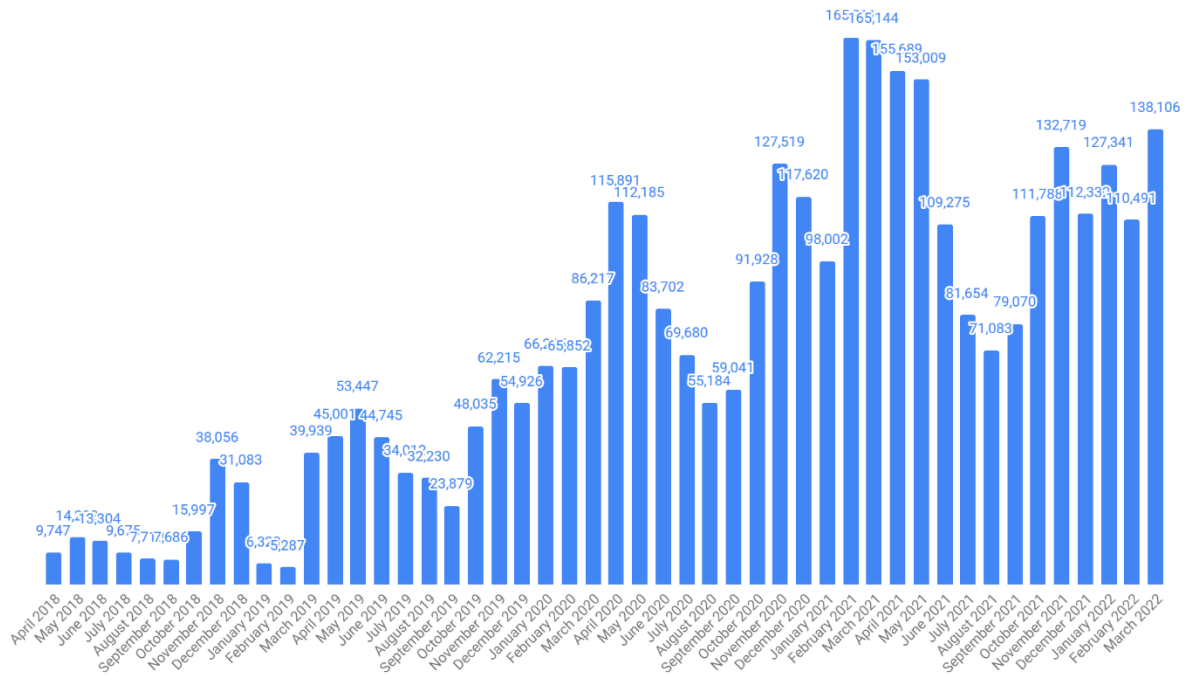


Figure 25: User-hours in each month since the project start.



3 Part 3: Platform for access to retrodigitised resources (LEX3)

This part of the infrastructure is covered with the task T8.3 which started in M12. The basic element of LEX3 platform is (legacy) Dictionary Viewer. As stated during the M18 review, the partner developing TEI Dictionary Viewer module (UT) reported severe delays in providing the software. The contingency plan is based on the possibility to use Lexonomy as the publishing platform also for legacy dictionaries available in TEI format.



4 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 BAŇSKI 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.

