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

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

Date: January 31st, 2019

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.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-INFRAIA-2016-2017
Grant Agreement No. 731015
ELEXIS - European Lexicographic Infrastructure

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

Deliverable Number: D8.1
Dissemination Level: Public
Delivery Date: January 31st, 2019
Version: 3
Author(s): Miloš Jakubíček, Ondřej Matuška,
Michal Cukr, Simon Krek, Michael Rundell

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
1, January 15th	Initial Draft	Matuška
2, January 20th	Added technical provisions	Jakubíček
3, January 27th	Assessment by AB member	Rundell

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Introduction

This report provides an assessment of the three parts of the ELEXIS infrastructure (LEX1, LEX2 and LEX3) during the first year of the project (M1-M12). Each of these parts is represented with its own task under the WP8 and this report is structured accordingly. Because the tasks related to LEX1 and LEX3 start not until M12, this report focuses mainly on the work that has been carried out in T8.2 on the LEX2 part of the ELEXIS infrastructure.

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.

Part 1: Lexicographic data integration platform (LEX1)

This part of the infrastructure is covered with the task T8.1 which starts in M12. The work done so far was largely focusing on the initial setup consisting of provisioning the necessary hardware and software equipment. 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 M12 the infrastructure consists of a server providing cloud services for hosting of all project data (corpora, dictionaries and tools part of LEX1).

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.

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

This part of the infrastructure is covered with the task T8.2 which started in M6. The LEX2 infrastructure is hosted and operated by Lexical Computing. In this report, we provide an overview of:

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

Tools and services in LEX2

access on **www.sketchengine.eu**

Sketch Engine is a 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 90+ languages and supports user corpus building in all of them. The largest corpora consist of texts in the total length of 35 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 variety of tools designed for searching effectively large text collections of billions of words according to complex and linguistically motivated queries. Sketch Engine is designed with a special emphasis on scalability and search speed.

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

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.

tokenization, part-of-speech tagging and lemmatization. Applying such tools manually requires programmatic and IT expertise which may represent an insurmountable obstacle for many researchers attempting corpus-oriented methods. Sketch Engine makes corpus building accessible to everybody with only general computer skills.

Word Sketch – The key tool in Sketch Engine which gave the system the 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 word sketch results is represented by Figure 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.

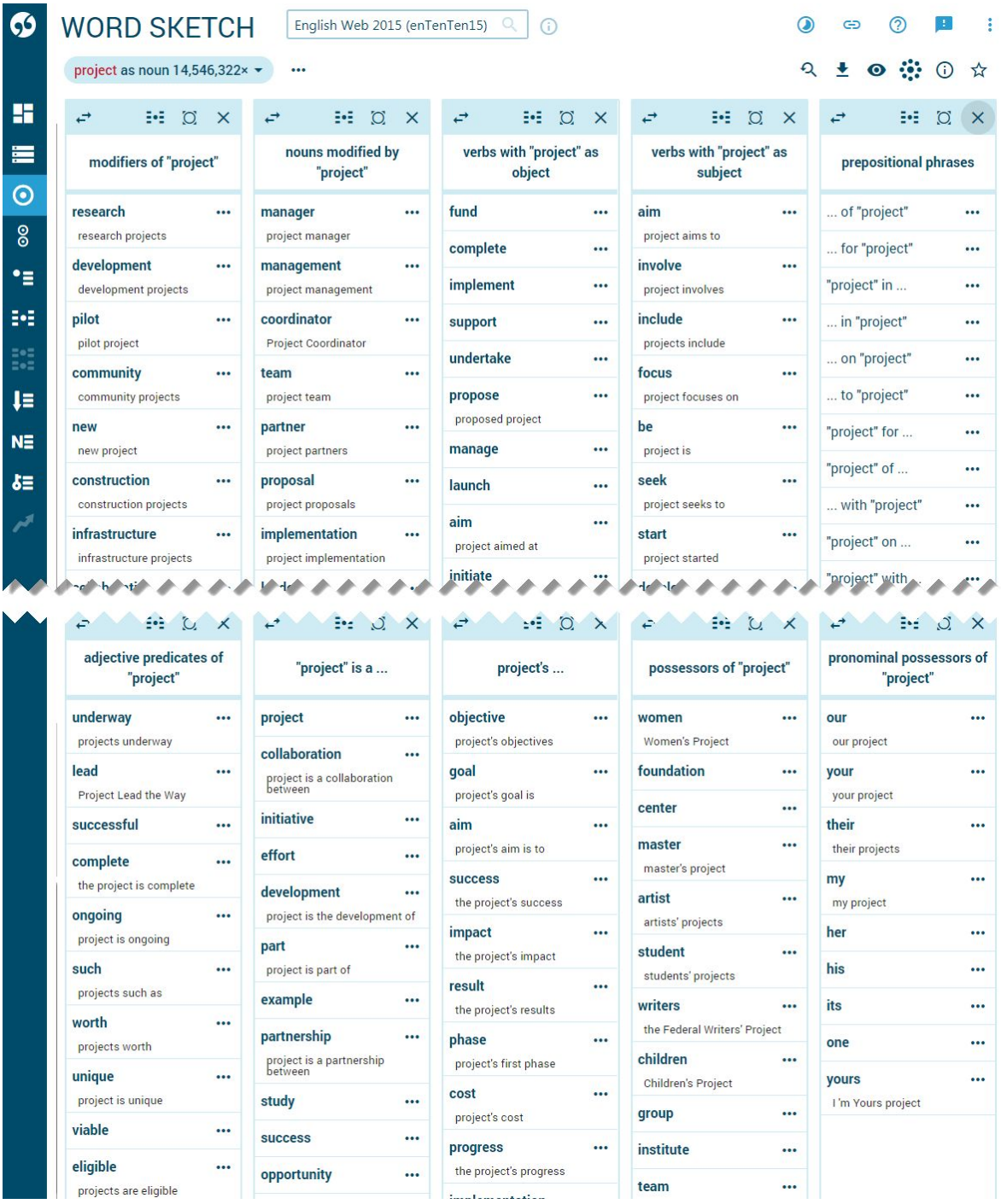

Figure 1. Results for the noun "project" from the Word sketch feature of Sketch Engine.

This visualisation (Figure 1) 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 14 million occurrences of the word *project*.

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

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.

to one word or the other. Figure 2 shows a layout of the tool. Find more information in [1]

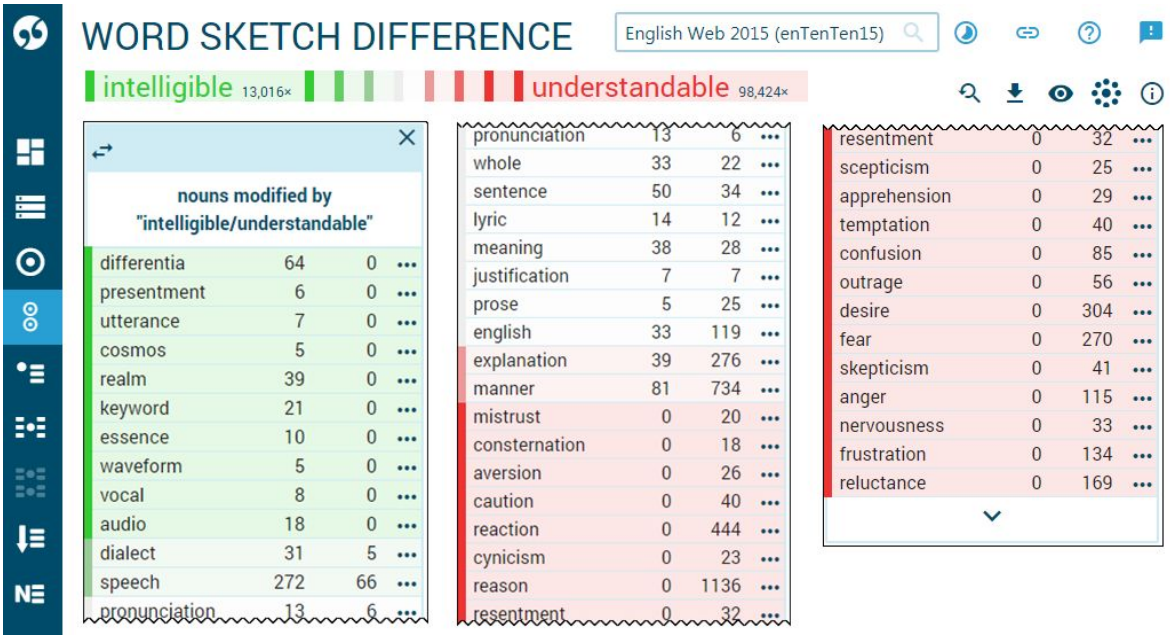

Figure 2. A layout of the word sketch difference tool for the adverbs “intelligible” and “understandable”.

Thesarurus – Sketch Engine 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 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]

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 3) 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.

It 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.

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.

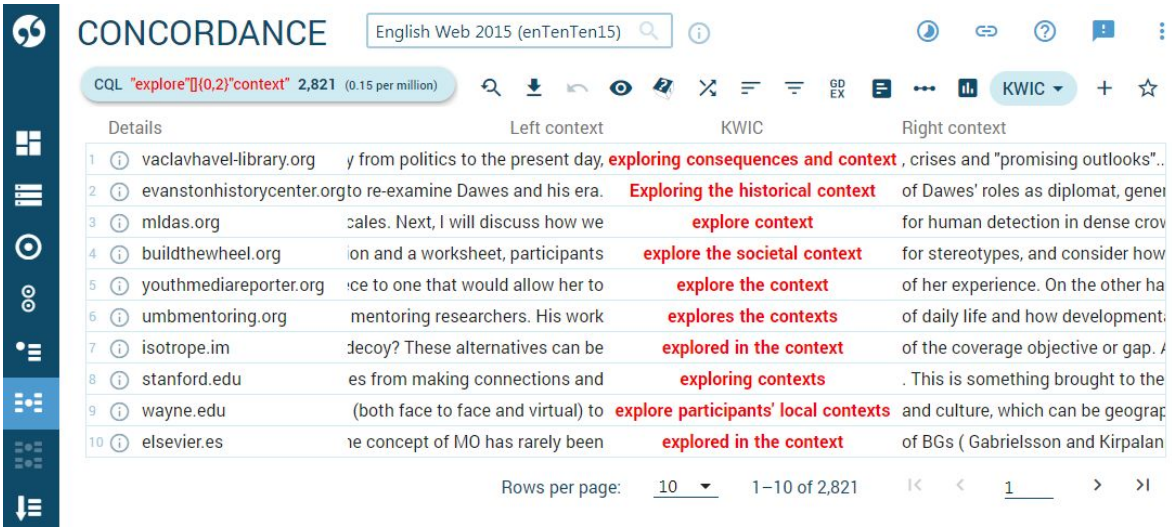

Figure 3. A concordance showing examples of a multiword query.

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.

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.

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.

OneClick Dictionary – The idea behind the OneClick Dictionary tool consists in the belief that dictionary making and dictionary editing could be much more productive, faster and cheaper if dictionary entries were pre-generated automatically with data coming from text corpora (Figure 4). Such dictionary drafts would still need to be post-edited by lexicographers but deleting, amending, 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.

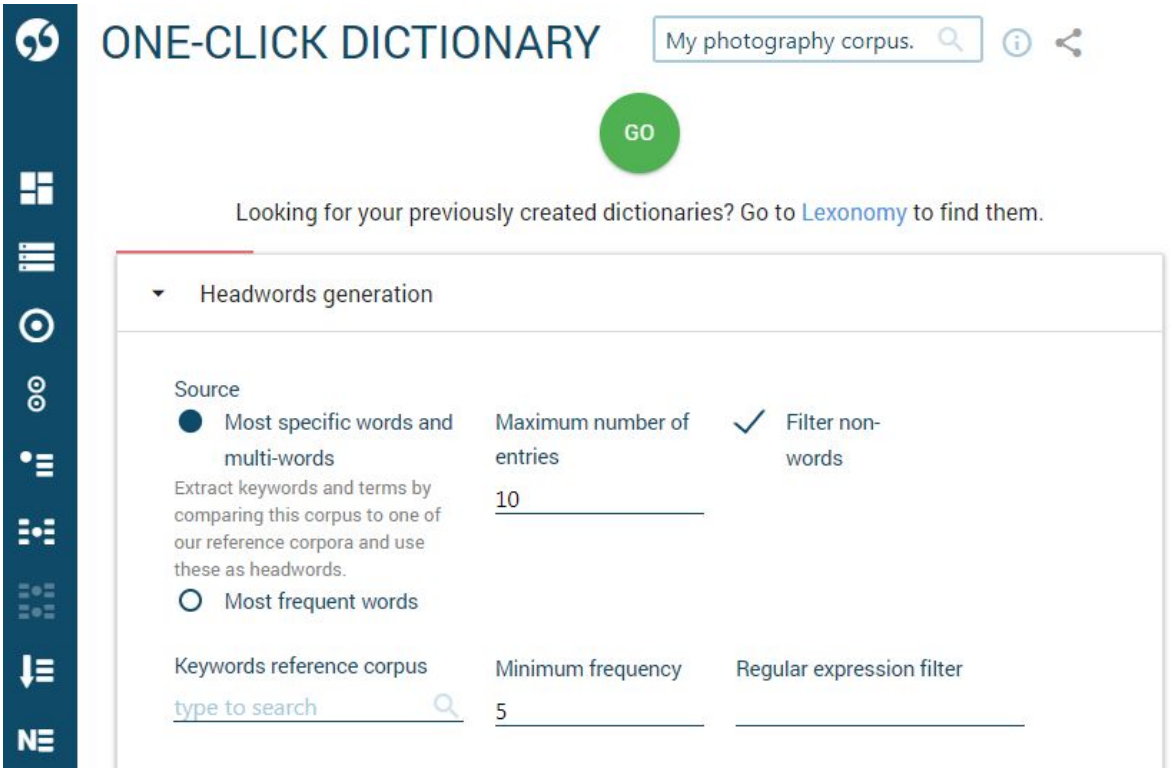

Figure 4. OneClick Dictionary – setting up the building of a new dictionary draft from a corpus.

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.

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.

access on **www.lexonomy.eu**

Lexonomy is a cloud-based open-source dictionary writing and online dictionary publishing system (see more in [3]) which is highly scalable and can adapt to large dictionary projects as well as small lexicographic works such as editing and online publishing of domain-specific glossaries, wordlists or terminology resources. Lexonomy allows editing from scratch but also accepts automatically generated dictionary drafts **pushed** to Lexonomy from Sketch Engine via a dedicated connection. During the editing process, users can also **pull** data from the corpora in Sketch Engine whenever they are needed during the entry editing process. The final dictionary can be exported or simply published online, accessible via a dedicated link in a desktop and mobile-friendly (Figure 10) user interface.

Figure 5. A dictionary entry within Lexonomy.

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.

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.

Dictionary templates

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

Editing the dictionary

The dictionary editing interface was specifically designed for users with little or no knowledge of the XML data format. [3] The interface automatically looks after the correct XML data structure (see Figure 6) and completely eliminates the error-prone procedure of typing the XML code manually.

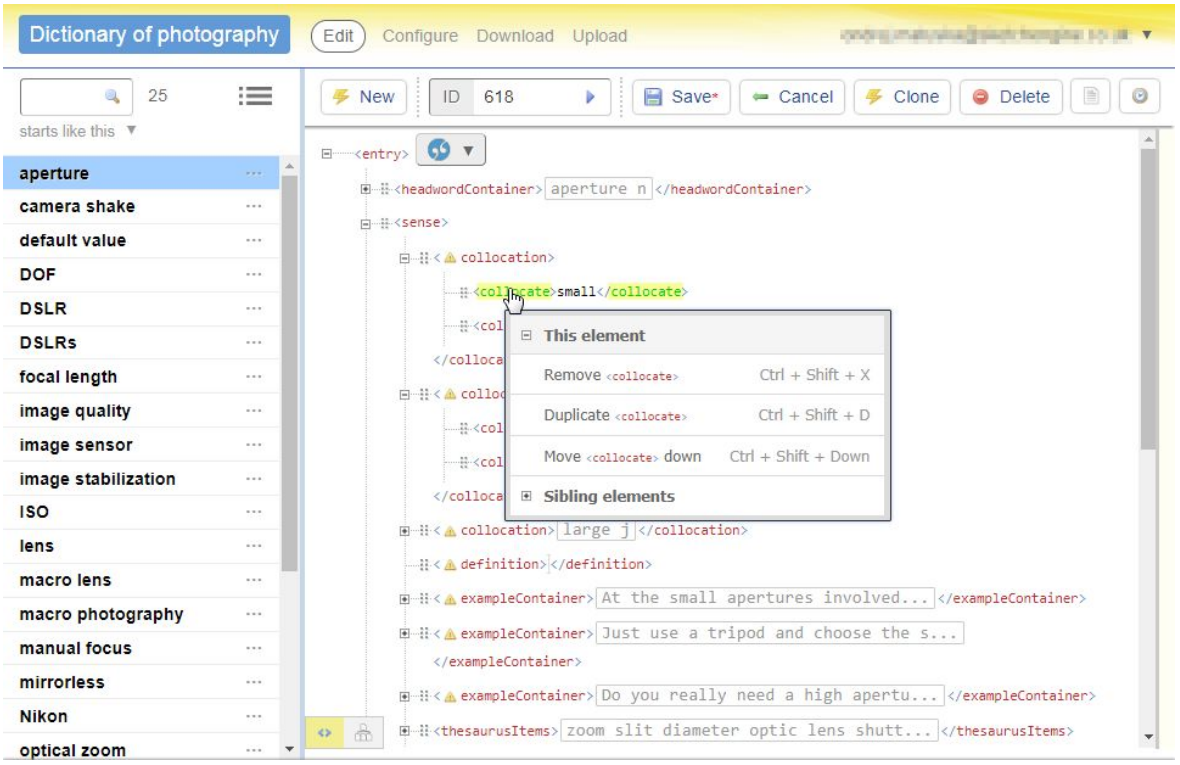

Figure 6. Editing particular attributes of a dictionary entry within Lexonomy.

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

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 7.

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.

Each dictionary project can be linked to a different corpus in Sketch Engine to acknowledge the fact that a domain specific glossary might need to draw data from a different data source (corpus) than a general language dictionary.

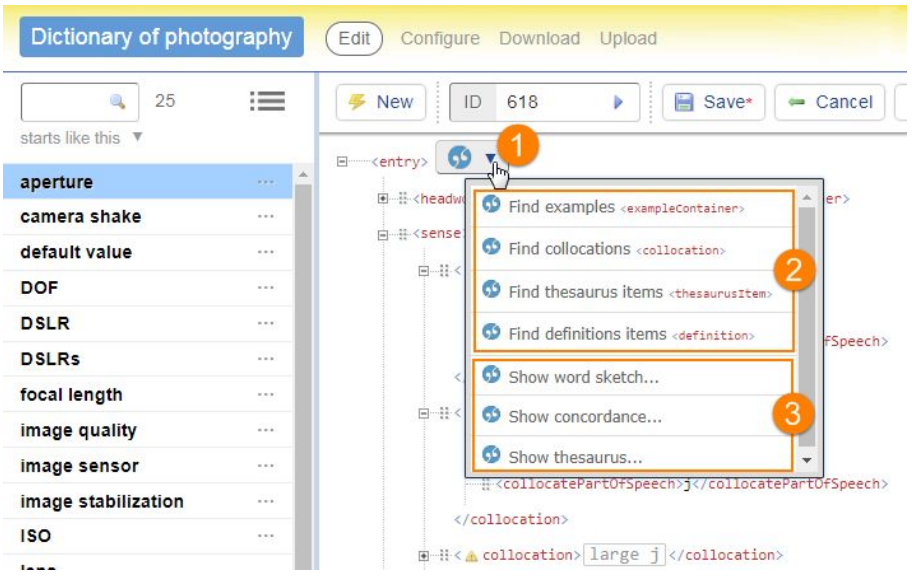

Figure 7. Interlinks between dictionary entries in Lexonomy and corresponding examples from Sketch Engine.

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.

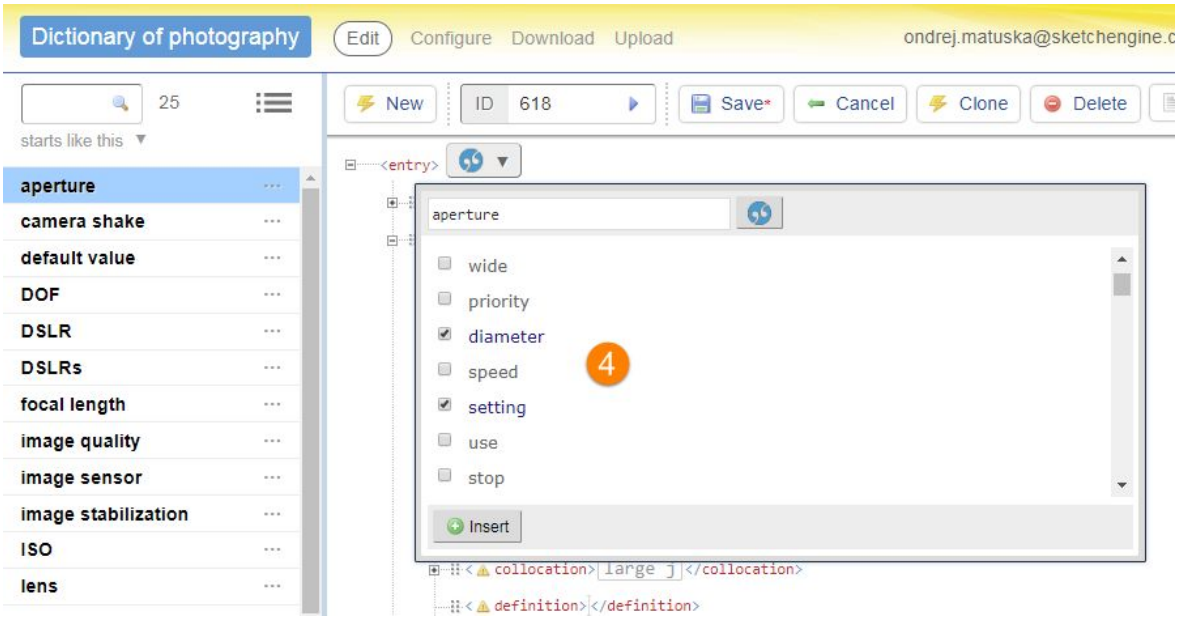

Figure 8. Lexonomy: data pulled from Sketch Engine can be edited.

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.

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.

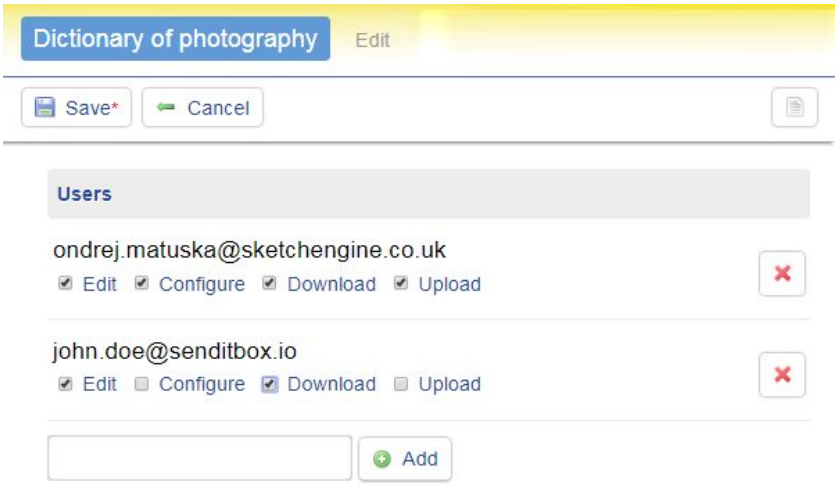

Figure 9. A list of access privileges to a dictionary in Lexonomy.

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

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.

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 10. The interface includes a 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.

Figure 10. 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.

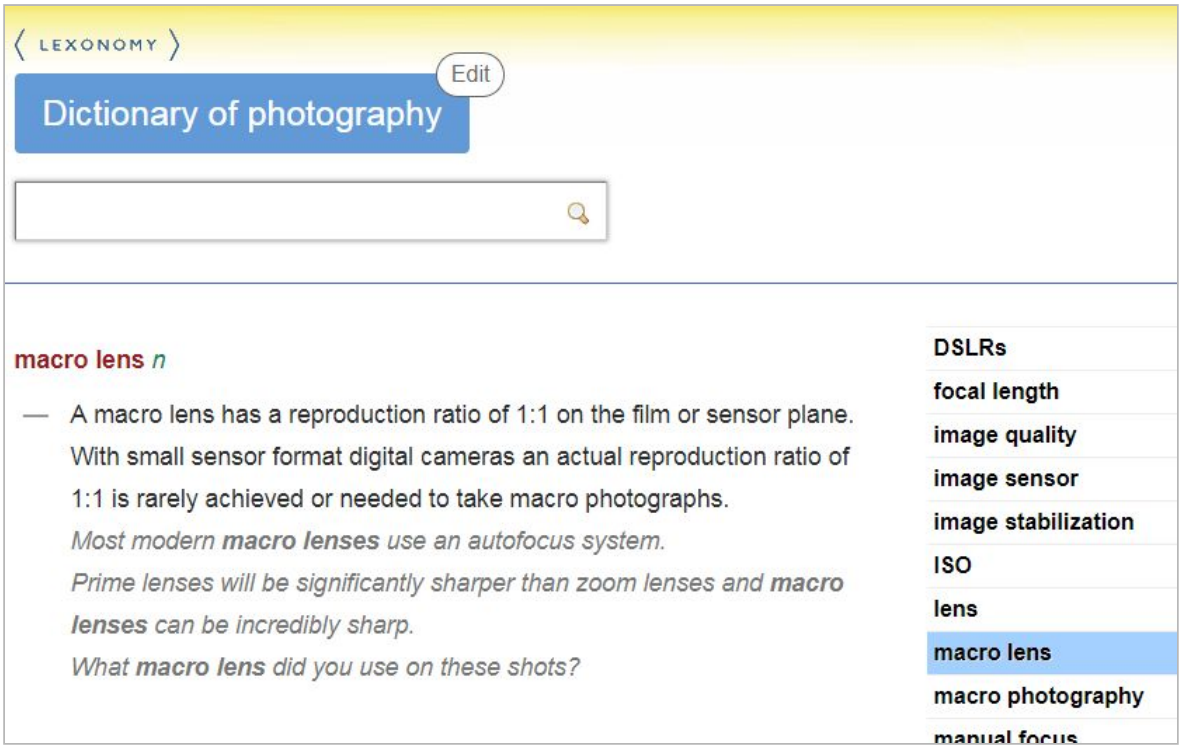

Figure 11. Lexonomy on desktop monitors.

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.

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 12**.** 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.

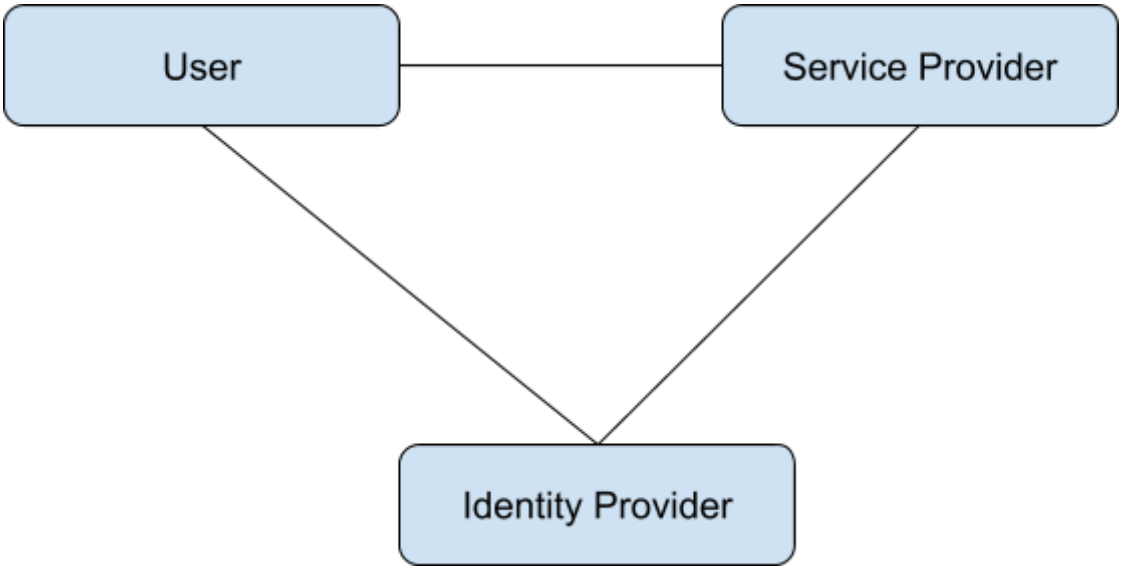

Figure 12. A simple diagram representing infrastructure of user access via SSO.

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].

[1] Detailed informationfor ELEXIS users are 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.

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.

Number of institutions

The Figure 13 gives overview of the number of institutions enjoying the access to the ELEXIS infrastructure.

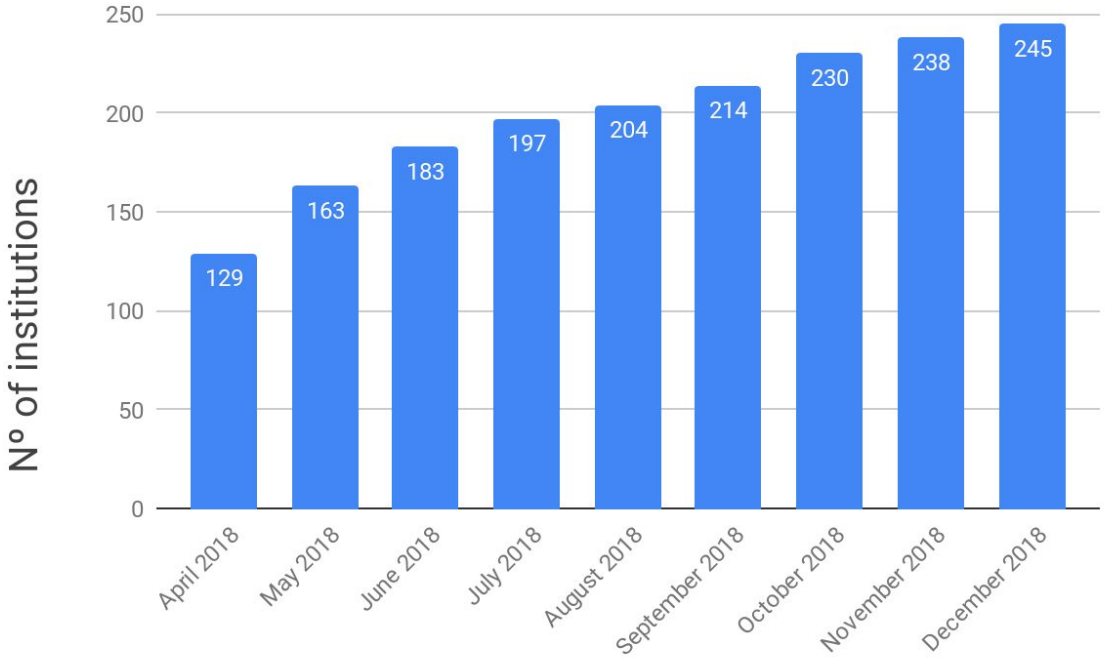

Figure 13. A total number of institutions joined the ELEXIS project per each month in 2018.

The growth was steeper at the beginning and by the end of the year, most key European institutions had gained access. A small number of institutions with outdated institutional systems are in the process of updating and configuring their systems to be compliant with the technical requirements for SSO access. More enrolment forms are still received every week and their access is set up typically within hours.

Country representation

Institutions from a total of 23 EU countries had shown interest in the ELEXIS infrastructure until the end of 2018 and submitted their enrolment forms (Figure 14)

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.

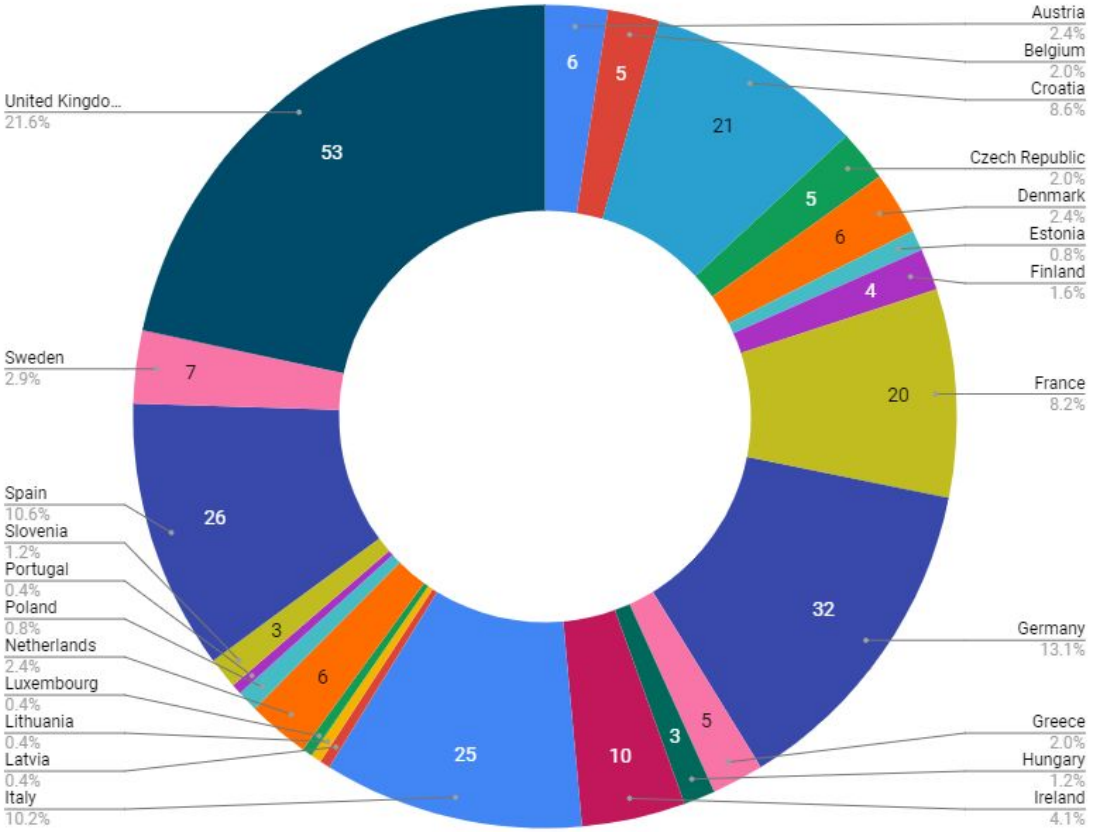

Figure 14. A number of institutions joined the ELEXIS project per country.

Two countries, Bulgaria and Slovakia, do not currently have a functioning national identity federation enabling their institution to gain access using the standard setup procedure. Lexical Computing is in touch with the institutions who are interested in gaining access and cooperates closely with their IT departments to provide access using alternative methods.

Number of user accounts

The number of user accounts saw its most prominent increase at the start of the winter semester in late September and early October. From just under 3,000 users at the end of the first month (April 2008), the infrastructure was used by very close to 10,000 users at the end of December 2018.

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.

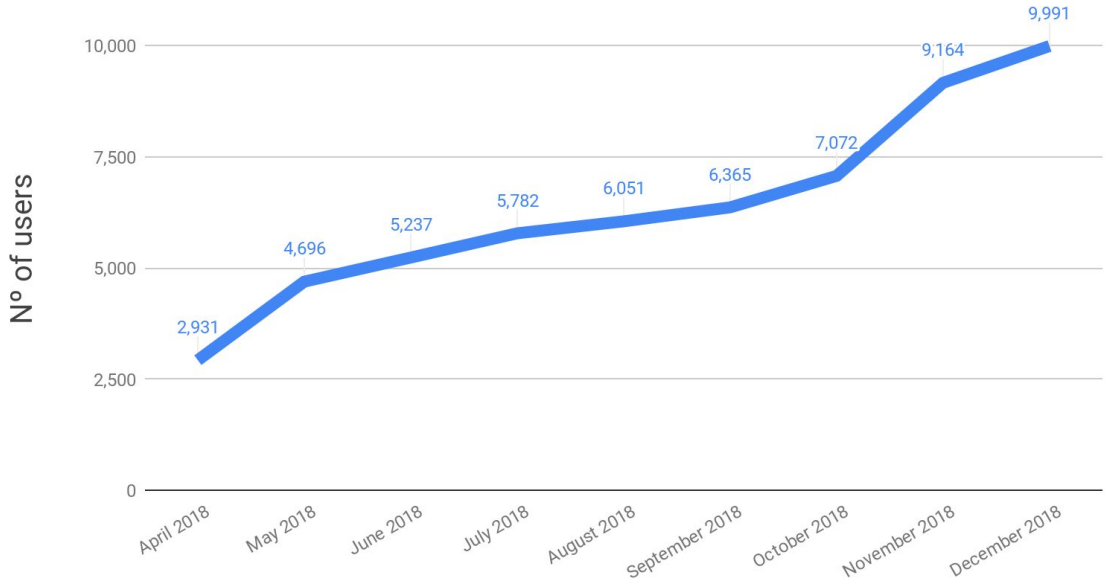

Figure 15. A total number of users gained access to Sketch Engine via the ELEXIS project (monthly view).

Average new user registration rate is 29 new user accounts every day including weekends and bank holidays.

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 generated some activity on the screen such as pressing the SEARCH button, changing the view settings, applying a filter etc.

In the period from April till December 2018, users taking advantage of the ELEXIS-funded access to the infrastructure generated a total of **147,490 user-hours**, see Figure 16.

The drops in the intensity of use correspond to the predictable events such as the summer holiday season (July, August), and Christmas holidays (December). The peaks correspond the end of the summer term when students finish their assignments (May, June), and when the winter term is in full swing (November).

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.

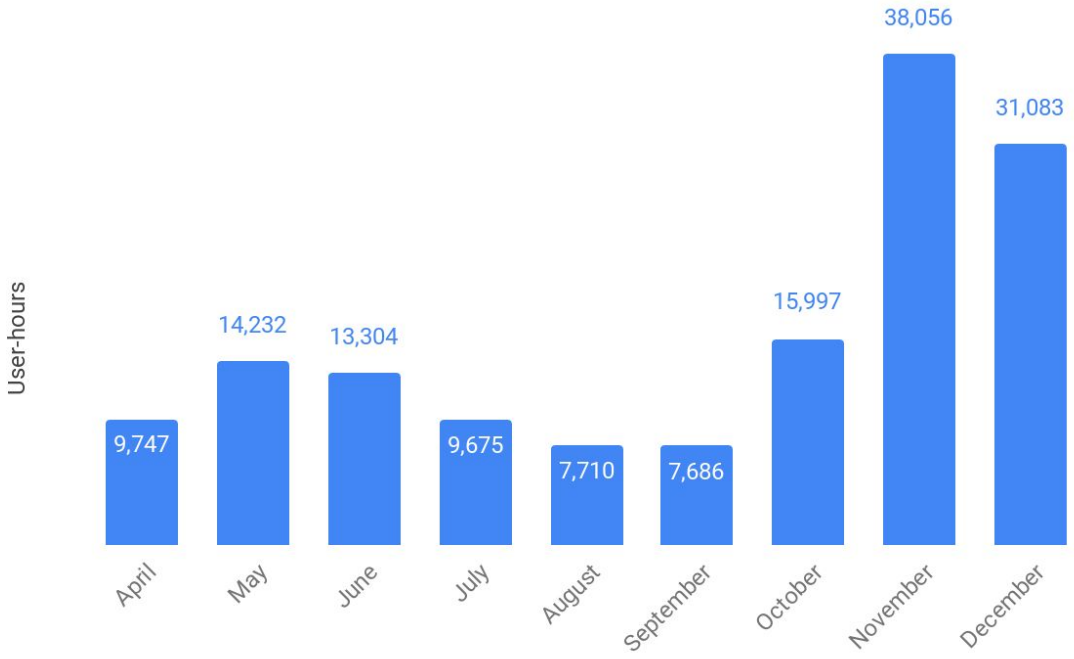

Figure 16. User-hours in each month (separately) during 2018.

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

This part of the infrastructure is covered with the task T8.3 which starts in M12. The work done so far was largely focusing on discussion of tools that will be initially part of the infrastructure. The LEX3 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.

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.

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.

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.

elexis european lexicographic infrastructure

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.