ELEXIFINDER:

A Tool for Searching Lexicographic Scientific Output

Iztok Kosem, Simon Krek

Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: iztok.kosem@ijs.si, simon.krek@ijs.si

Abstract

Access to lexicographic research is highly important for lexicographers when conceptualizing and compiling dictionaries, and preparing their publications for presentation to the lexicographic community. There have been several attempts to offer a systematic record of lexicographic scientific output, and advanced search of it, but most of them are no longer updated, focus only on bibliographic data, and do not include works from other fields related to lexicography. The tool called Elexifinder has been developed within the European Infrastructure for Lexicography (ELEXIS) project in order to facilitate knowledge exchange in the lexicographic community and promote open access culture in lexicographic research. In this paper, we present the first version of the tool that contains 1,755 publications and 78 videos in 11 different languages, and offers various search options to users. We describe the Elexifinder architecture, the process of including content, and present the interface's features. The paper concludes with the presentation of future plans, including the various publications that will be included in the next version of the tool.

Keywords: Elexifinder; lexicographic research; ELEXIS; lexicography; online tool

1. Introduction

In state-of-the-art lexicography, it is paramount that lexicographers have access to resources such as corpora and other dictionaries, and tools such as dictionary-writing systems and corpus query systems. Yet, it is equally important that lexicographers have constant access to scientific output in lexicography and disciplines related to lexicography, so they can follow the projects and research of their colleagues around the world, develop new ideas, conceptualize dictionaries, understand and address linguistic problems, and position their own work in the lexicographic community.

One of the issues faced by lexicographers is that lexicographically-relevant scientific output is very scattered. Journals focused on lexicography (the *International Journal of Lexicography*, *Dictionaries*, *Lexicographica*, *Lexikos* etc.) are published by different publishers. Then, each lexicographic association has its own proceedings, and moreover, their availability varies – some associations have all their proceedings freely available on their website (e.g. EURALEX), while others only the more recent ones (e.g. ASIALEX). Accessibility is especially an issue with older literature and books, as in most cases these resources are available only in print (although many of these publications have been digitized and can be searched in Google Books).

Further difficulty in finding lexicographically-relevant scientific output lies in the fact that many fields are of relevance to lexicography, for example lexical semantics, pragmatics, corpus linguistics, and more recently natural language processing. This means that lexicographers need to constantly follow journals, proceedings and other resources covering those fields for any relevant papers.

An additional obstacle to this form of knowledge exchange is language. Namely, lexicographers are usually very familiar with lexicographic research in their native language, and possibly in other languages they are fluent in. They can also fairly easily find lexicographic research in English, not only because there is an abundance of literature available but also because it is much better covered by search engines. However, to identify the relevant literature or authors in other languages is much more difficult. This can lead to isolation of researchers or communities, especially the ones that do not (also) publish in English. Thus, their work, as relevant and innovative as it may be, stays unnoticed in other communities.

In this paper, we present the Elexifinder tool which addresses these issues and has been developed within the European Infrastructure for Lexicography (ELEXIS))[1], a H2020 project funded by the European Commission. First we conduct an overview of some existing efforts in collecting and recording lexicographic research, and their relevance for the development of the tool. This is followed by the description of the tool, the technology behind it and the procedure for including research papers. Next, we present parts of the tool interface and current contents. We conclude by discussing a few potential use cases and presenting plans for the future, both in terms of content and features.

2. Past and current efforts

There have been several known attempts to make an inventory of lexicographic literature, which have been focussed on collecting research publications, bibliographic information on publications and/or dictionaries, or both. We make an overview of them in this section.

Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung by Wiegand (2012) is a five-volume bibliography that covers the largest number (33,339) of lexicographic works of all resources listed here. The shortcomings of this resource are that it is available in print only and that it is mainly focused on the field of German studies. Similarly limited in scope is Ahumada's (2016) collection of 6,560 items mainly from Hispanic (meta)lexicography.

[1] https://elex.is/

The second largest bibliography of lexicography is the one by Córdoba Rodríguez[2], which contains 10,192 items published between 1940 and 2003, including relevant newspaper articles. The search can be conducted by thematic blocks or by authors (alphabetically). One shortcoming of this bibliography is that it is no longer updated.

Also in need of an update is the International Bibliography of Lexicography, initiated by the European Association for Lexicography (EURALEX).[3] It contains approx. 2,000 entries[4] that can be viewed thematically or alphabetically. It also includes links to lists of reference portals and lexicographic bibliography collected by R.R.K. Hartmann (2007). As it is stated on the resource website, the bibliography has not been updated since 2012. In addition, the website offers rather limited search options.

The Online Bibliography of Electronic Lexicography (OBELEX) is an ongoing project at the Institute for the German Language (Möhrs & Töpel, 2011), which has started in 2008 and consists of two databases. The first database, called OBELEXdict includes over 17,000 online dictionaries (Möhrs, 2016), but as the authors point out on the resource website,[5] "the term 'dictionary' […] has a broad interpretation, i.e. all word-related reference works were included, without the quality of the content having been checked". The users can search the database by type of dictionary, title, language (family), limit the search to dictionaries with audio or video files, illustrations, etc. The second database, called OBELEXmeta,[6] contains bibliographic information on around 2,000 entries, which cover articles, monographs, anthologies and reviews. Most of the works in the database have been published from 2000 onwards, but some older relevant works are also included. Advanced search options, such as searching by title, author, year of publication, language and keywords, are provided.

A more recent large-scale bibliographic project proposal is LexBib by Lindemann et al. (2018) that aims to create

> "a domain-specific online bibliography of lexicography and dictionary research (i.e. metalexicography) which offers hand-validated publication metadata as they are needed for citations, and which in addition is complemented with the output of an NLP toolchain." (ibid: 699)

In addition to ensuring a comprehensive coverage of lexicographic literature,[7] the importance of this proposal lies in enabling easy citation extraction and the introduction of automatic keyword indexation (and evaluation). In the first, testing

[2] Accessible at http://www.udc.es/grupos/lexicografia/bibliografia/index.html.

[3] Accessible at http://euralex.pbworks.com/w/page/7230036/FrontPage.

[4] The exact number is not provided on the resource website.

[5] https://www.owid.de/obelex/dict/en?info

[6] https://www.owid.de/obelex/meta/en

[7] https://www.zotero.org/groups/1892855/lexbib/items

phase, the authors propose including only items in English, published between 2000 and 2017.

One of the advantages of LexBib is that the focus is not only on recording bibliographical data but also on indexing full-text publications. And when talking about large collections of full-text lexicographic publication, we must definitely mention an impressive lexicographic corpus of over 5,000 lexicographic articles and books (29.2 million tokens), compiled by Gilles-Maurice de Schryver and used in studies such as Lew and de Schryver (2014) and, part of it, de Schryver (2009, 2012). The corpus is not publicly available but all efforts should be made to make further use of all the manual labour that has been put into preparing the texts of this corpus.

We can conclude that existing resources on lexicographic research are mainly focused on bibliographical aspects, especially the information about titles, authors, and keywords. Furthermore, some of the resources are no longer updated or have limited coverage. An additional problem is accessibility, as certain resources are available only in print or are private collections. One thing that none of the existing or planned resources address is the fact that scientific output is no longer limited to articles and books. It has become multimodal; there are now many video presentations on important lexicographic topics available.

3. ELEXIFINDER

The European Infrastructure for Lexicography (ELEXIS) is a project running from 2018-2022, with the aim to build a sustainable infrastructure for lexicography (cf. Krek et al., 2018). The objectives emphasized in ELEXIS are the following: the infrastructure will (1) foster cooperation and knowledge exchange between different research communities in lexicography in order to bridge the gap between lesser-resourced languages and those with advanced e-lexicographic experience; (2) establish common standards and solutions for the development of lexicographic resources; (3) develop strategies, tools and standards for extracting, structuring and linking of lexicographic resources; (4) enable access to standards, methods, lexicographic data and tools for scientific communities, industries and other stakeholders; (5) and promote an open access culture in lexicography, in line with the European Commission recommendation on access to and preservation of scientific information.

Fostering knowledge exchange in lexicography is thus one of the main objectives of ELEXIS, and improving access to lexicographic scientific output falls very much under this description. This solution will be provided in the form of a tool called Elexifinder[8] that aims to become some sort of lexicographic Google and will not only help lexicographers in finding the relevant literature, but also allow contributions of papers or suggestions for further inclusion in the tool. The tool also addresses another objective

[8] http://er.elex.is

of the project, namely promoting open access culture, as open access publications are given priority in the inclusion process.

3.1 Elexifinder architecture

Elexifinder has been built using some of the elements of the Event Registry system architecture (Leban et al., 2014; Leban et al., 2016a). Event Registry[9] is a system used for identifying world events from news articles. Articles in different languages are collected as soon as they are detected by the Newsfeed service, then semantically enriched, and clustered to detect events (i.e. articles covering the same event). Semantic enrichment includes the identification and disambiguation of so-called concepts, which include named entities (people, places, locations) as well as non-entities or topics. Concepts are identified by wikification, "a process of entity linking that uses Wikipedia as the knowledge base" (Leban et al., 2016b).

Elexifinder uses only a portion of this system, namely the semantic enrichment (to enable various search options) and the interface. For the first version, the decision was made not to make significant modifications to the interface, as we wanted to have some content to be able to properly evaluate the usefulness of its functionalities.

3.2 Data collection and preparation

Preparing a publication for insertion into Elexifinder consists of two steps: the preparation of metatextual information and the preparation of publication content. The following metatextual information is recorded:

- Publication title
- Publication authors[10]
- Publication keywords (if available)
- Publication source. This can be the name of a conference or a journal, usually the year and/or the number of the issue is also included, e.g. EURALEX 2016 or Lexikos 2013-13.

[9] http://eventregistry.org/

[10] One of the issues that has been identified only after the launch of the first version of Elexifinder was multiple variants of authors' names, such as John Sinclair and John McHardy Sinclair, or Danie Prinsloo, Danie J. Prinsloo and D.J. Prinsloo. This will be corrected for the second version of the tool by establishing the links between these variants and choosing one of them as the canonical form for Elexifinder. Slightly problematic will be authors that have changed surnames, e.g. Annette Klosa and Annette Kückelhaus, as both forms can actually be considered canonical.

- Publication language. ISO3 language codes are used.
- Publication URL. If the publication is not available online (e.g. in case of a book), the link points to the publisher website where the book is presented.
- Publication date. The format used is YYYY-MM-DDThh:mm:ss. For conference proceedings, the first day of the conference is used if no other date is provided. For journal issues, the last day of the issue scope is used, for example if the journal has four issues per year, the date used for the first one is at the end of the first quarter (i.e. the end of March).
- Location of the source. Recorded as a URI (Uniform Resource Identifier) of the city of the publication publisher or conference, which can be found using the Autosuggest location service by the Event Registry (http://eventregistry.org/documentation?tab=suggLocations).
- Location of the first author. Recorded as a URI of the city (of the affiliation) of the first author.

At the moment, the metatextual information is recorded in an Excel spreadsheet. This has the advantage of easy copying of repeating information such as publication name or location URI. It is planned to later offer an online form for individual contributions where the metadata entry would be even easier.

The second part is content preparation. Publications are usually obtained in PDF format, although if DOC(X) format is available it is preferred. The first step consists of converting the files into the TXT format, followed by checking the files and correcting any conversion errors. At this point, a copy of TXT versions – which at this point still reflect the PDF originals – is archived. This is to ensure that they can be used for any (corpus) analyses that require entire texts. The next step is the removal of content not needed for semantic enrichment: header and footer information (often repeated on every page), page numbers, publication title, author information, abstract, keywords, and references. In addition, figures, tables (and titles), footnotes and appendices are often removed, although in the case of tables that often depends on their content. For example, tables containing (only) statistics are removed but tables containing textual information are not. Similarly, footnotes containing only URLs are removed but footnotes containing remarks are not. In general, the focus is on maintaining the content that can be most informative about the topic(s) of the publication.

We have also decided to include videos of presentations with lexicographically relevant content. The same metatextual information is recorded, and the content used for semantic enrichment is an abstract or accompanying text (e.g. an abstract of the presentation at the conference).

Both metatextual information and content are then transferred into a JSON file which is needed for Elexifinder preparation.

3.3 Current contents

At the time of writing, Elexifinder included 1,755 publications and 78 videos in 11 different languages. The contents were:

- EURALEX conference proceedings from 1983 to 2016 (1,552 papers in total).
- eLex conference proceedings from 2009 to 2017 (203 papers in total)
- 21 video presentations from the eLex 2011 conference
- 33 video presentation from EURALEX 2018 conference
- 18 video presentations from various symposia in Slovenia (seven in English, 11 in Slovene)
- six video presentations from the WNLEX Workshop 2018 in Ljubljana

Importantly, all the content in Elexifinder at the moment is open access. Open access publication will continue to be prioritized, and when we start including books and monographs later we will make an appeal to publishers to publish the PDF versions of the publications, especially older ones, somewhere on their website.

3.4 Elexifinder interface

In this section we present the Elexifinder interface, including various search options available to the users. Elexifinder consists of a search window, filter line and result window. The search window offers users the option to search by keywords, either in publication title or body, or by concepts (named entities or topics). The auto-suggest functionality facilitates searching (see Figure 1). In addition, advanced search commands are supported, for example by using a – sign before a keyword one can limit the search results not to include a certain keyword (e.g. dictionary –thesaurus returns publications with a keyword “dictionary” but not including “thesaurus”).

The line with filters enables filtering by Locations (of the first author), Sources (journal or conference, and/or a specific author), Category, Time of interest (from/to specific date or period), Language, and data type (text or video). These filters can be used independently or in combination with the keywords in the search window. For example, by leaving the search window empty and using the Locations filter, one can search for all the publications coming from authors from a certain country.

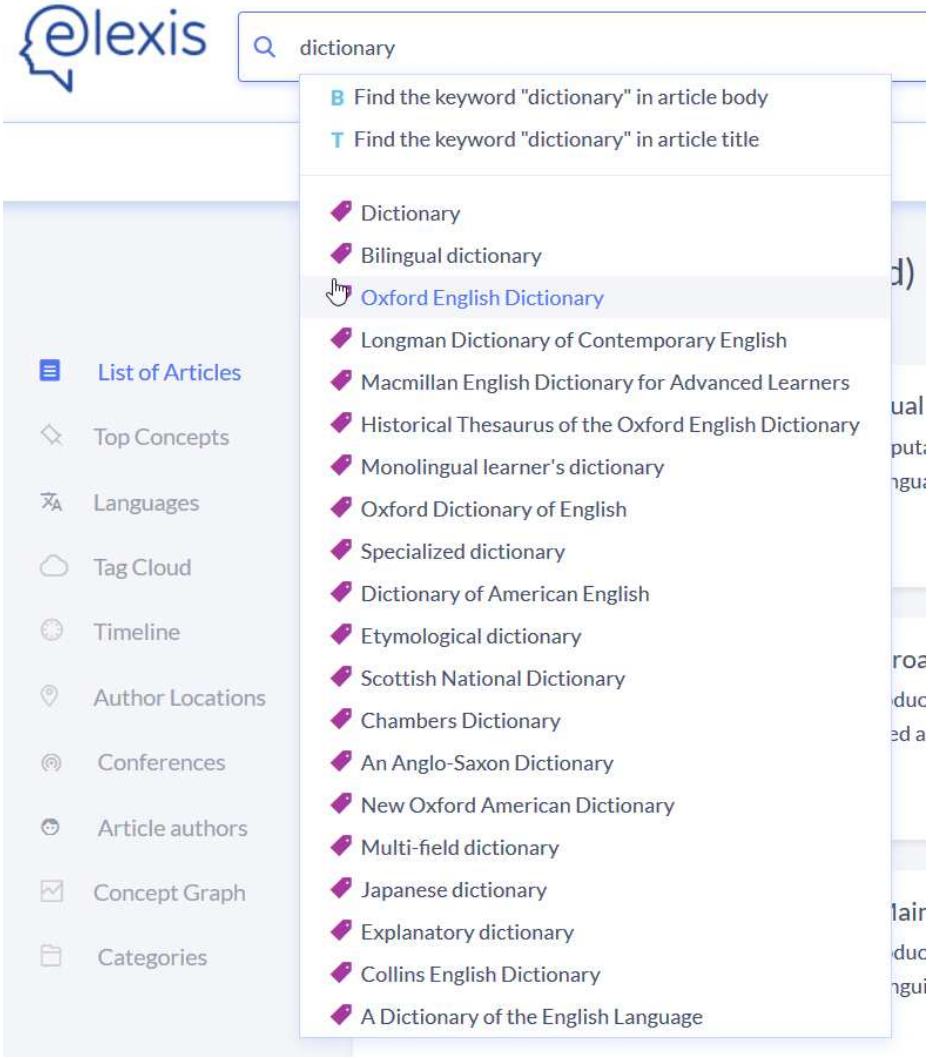

Figure 1: Auto-suggest functionality in search

The results part of the Elexifinder interface is dedicated to showing the search results. On the homepage, before any search is conducted, a map with the locations of all the authors of all the publications in Elexifinder is shown by default (Figure 2). Once any search is conducted, the results window offers a list of publications found (right-hand panel) and a left-hand menu with different visualization options. For each result in the list, the information on title, author(s), source and date of publication is provided, and in the default List view, first few lines of the text are also shown. The other types of view are Grid (each type of information is clearly named), Compact (List view but without the first few lines of text), and Details (List view + a list of most relevant semantic concepts).

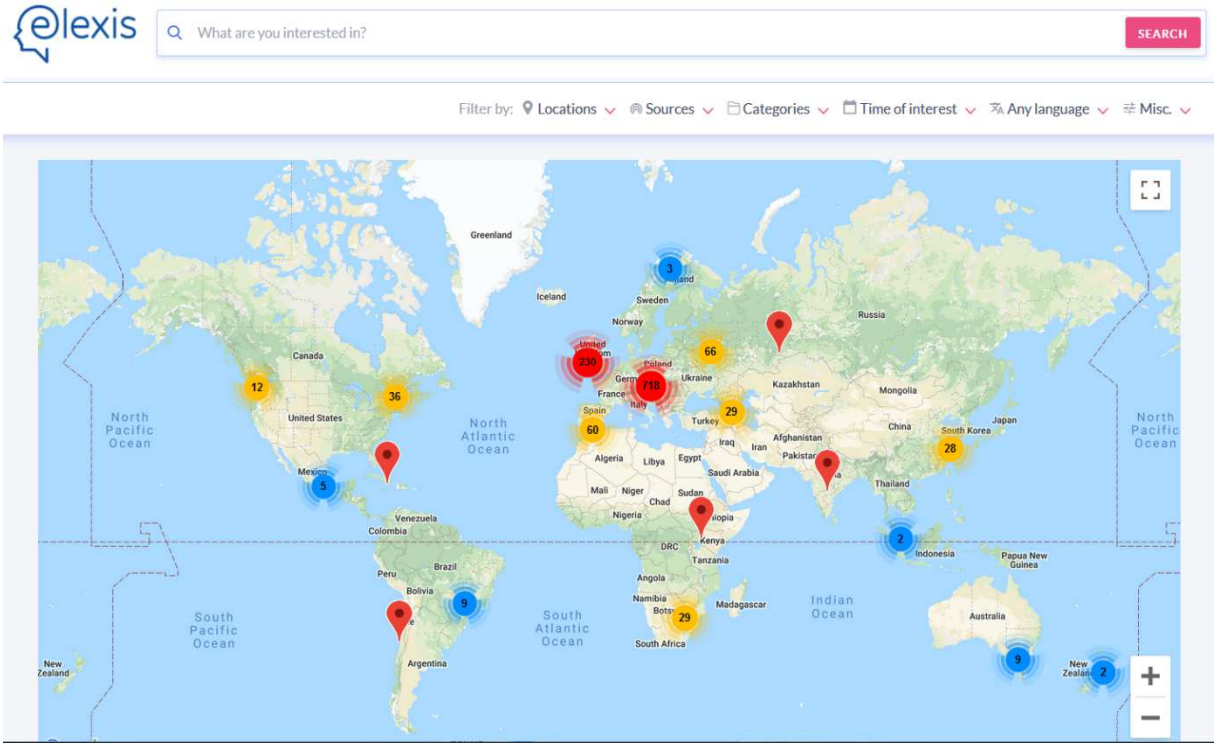

Figure 2: Homepage of Elexifinder

In the left-hand menu, the users can obtain more information on the results, and filter them (by clicking directly on maps or diagrams). Available options:

- **Top Concepts** provides a list of most relevant concepts found in the results. Concepts can be listed by relevance (default setting), frequency or uniqueness. By clicking on any concept, it is possible to limit the search further (in addition to the search condition).
- **Languages** displays a list of languages in which the publications in the results are written.
- **Tag Cloud** of top keywords in the results. Any keyword is clickable to further limit the search results.
- **Timeline** shows a distribution of results on a timeline. Daily, weekly or monthly view is available.
- **Author Locations** offers a map of locations of the first authors, and a diagram showing the number of publications per country (based on first author information).

- **Conferences**[11] includes three features: a diagram of all the conferences, with the number of publications; a map with locations of sources and number of publications, and a diagram showing numbers of publications per country.

- **Article Authors** offers a list of authors (not only first authors) that have authored the most publications in the results.

- **Concept Graph** shows a graph of most frequent concepts, with links between them if they exist.

- **Categories** is a visualization of automatically assigned categories and subcategories to the results. At this moment, these are still general categories (taken from DMOZ[12]) rather than categories adapted to lexicography.

A useful feature is the option to download an image of every diagram, map or cloud shown in Elexifinder. Moreover, certain features such as Top Concepts also offer the option to download the data displayed in the diagram in the TSV format.

4. Future plans

Elexifinder was launched at the beginning of 2019, and an extensive list of publications and video recordings for further inclusion has already been prepared. This includes journals such as *Dictionaries*, *IJL*, *Lexikos*, *Lexicon*, *Lexicographica*, *Nordiske Studier i Leksikografi*, *Slovenščina 2.0* and others, and proceedings of Asialex, LexicoNordica, GLOBALEX workshops, etc. Also on the list are collective volumes, monographs and similar works. As far as videos are concerned, we aim to include video presentations from all the relevant conferences (e.g. eLex, EURALEX, Asialex) and other specific international or national events. Moreover, videos of interviews with (famous) lexicographers (e.g. the FutureLearns interview with Michael Rundell) [13] will be included.

It is important to note that many items mentioned above will likely not be collected anew, but will be obtained from Gilles-Maurice de Schryver and the LexBib team. This will prevent the duplication of effort and enable the focus of further work on missing content, i.e. content currently not covered by any of the existing bibliographic or textual resources. This is particularly the case with publications that are not in English.

Also, special attention needs to be paid to research works in non-lexicographically dominated publications. As already mentioned in the beginning, there are many fields that produce research relevant for lexicographers, and tracking down such papers can

[11] This feature will be renamed after journal papers and other publications are added.

[12] https://en.wikipedia.org/wiki/DMOZ

[13] https://www.youtube.com/watch?v=5NO2YfJlXOA

be challenging. The first obvious step would be to track down lexicographically-related special issues, but much more work is needed to identify individual papers. To improve the coverage of Elexifinder and to ensure quick updating of its database, it is envisaged that members of the lexicographic community will be able to directly contribute to the resource, either by suggesting relevant publications or videos for inclusion, or by providing the content and its metatextual information directly. Besides the obvious benefits of recently published works being immediately available to the community, there are benefits for editors, reviewers and other people involved in publication preparation as they will be able to search for any related publications of the same author(s) with the same or similar content.

In addition to enhancing the Elexifinder database with new content, improvements of the frontend are planned. The first part of the improvements is connected to searching. This includes cross-lingual searching, which would enable automatic translation of search terms into all other languages of publications found in Elexifinder. Such a feature is already part of the Event Registry system and its identification of events, so the aim is to adapt it to the needs of Elexifinder.

Partly linked to cross-lingual searching is the introduction of a new and more lexicographically-oriented categorization of publications, or freshly devised "ontology for lexicography", which would replace the existing DMOZ-based categorization. For this, we will work together with the LexBib team on keyword indexation (and evaluation) in order to devise a common solution. Also, existing ontologies such as the META-SHARE ontology (McCrae et al., 2015) will be used as a starting point.

Other improvements will be done on the homepage, where we plan to include additional content that would help promote lexicography and attract users to the website more regularly. Such content includes the list of most searched/clicked publications or videos, alerting users to the most recent inclusions, presenting a list of publications (in different languages) on a certain topic of interest, listing conference and journal calls, etc.

In sum, Elexifinder will become an integral part of ELEXIS infrastructure that will be complementary to other resources and tools developed within ELEXIS, and continuously improved long-term with the help of the lexicographic community.

5. Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

6. References

Ahumada, I. (2016). Metalexicografía del español: clasificación orgánica y tipología de los diccionarios en el Diccionario Bibliográfico de la Metalexicografía del Español (DBME). In *Anuario de estudios filológicos*, (39), pp. 5–24.

De Schryver, G.-M. (2009). Bibliometrics in Lexicography, *International Journal of Lexicography*, (22,4), pp. 423–465.

De Schryver, G.-M. (2012). Trends in Twenty-Five Years of Academic Lexicography. *International Journal of Lexicography*, (25,4), pp. 464–506.

Euralex: International Bibliography of Lexicography. Accessed at: http://euralex.pbworks.com. Date of access: 15th May 2019.

Hartmann, R.R.K. (2007). *Bibliography of Lexicography.* Accessed at: http://euralex.pbworks.com. Date of access: 15th May 2019.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 881-891. Available at: http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2986-1-10-20180820.pdf.

Leban, G., Fortuna, B. & Grobelnik, M. (2016a). Event Extraction from Media Texts. In C. Sammut & G. I. Webb (eds.) *Encyclopedia of Machine Learning and Data Mining.* Boston: Springer, pp. 1-7. Available at: https://link.springer.com/content/pdf/10.1007%2F978-1-4899-7502-7_901-1.pdf.

Leban, G., Fortuna, B. & Grobelnik, M. (2016b). Using news articles for real-time cross-lingual event detection and filtering. *First International Workshop on Recent Trends in News Information Retrieval (NewsIR'16), Padova, Italy.* Available at: https://pdfs.semanticscholar.org/f917/c0cff24fed1af45f94c53b74ca0229874966.pdf.

Leban, G., Fortuna, B., Brank, J. & Grobelnik, M. (2014). Event Registry: Learning About World Events from News. *Proceedings of the 23rd International Conference on World Wide Web*, pp. 107-110.

Lew, R. & de Schryver, G.-M. (2014). Dictionary Users in the Digital Revolution, *International Journal of Lexicography*, 27(4), pp. 341–359. Available at: https://doi.org/10.1093/ijl/ecu011.

Lindemann, D., Kliche, F. & Heid, U. (2018). LexBib: A Corpus and Bibliography of Metalexicographical Publications. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 699-711. Available at: http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2929-1-10-20180820.pdf.

McCrae, J., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., Cimiano, P. (2015). One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. In: *Proceedings of 12th Extended Semantic Web Conference (ESWC 2015).* Portorož, Slovenia. DOI: 10.13140/RG.2.1.3233.6244

Möhrs, C. & Töpel, A. (2011). The "Online Bibliography of Electronic Lexicography" (OBELEX). In: I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, 10 - 12 November 2011.* Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 199-202.
Available at: http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-26.pdf.

Möhrs, C. (2016). Online Bibliography of Electronic Lexicography. The Project OBELEXmeta. In T. Margalitadze & G. Meladze (eds.) *Proceedings of the 17th EURALEX International Congress: Lexicography and Linguistic Diversity. Tbilisi, Georgia 6-10 September 2016*, pp. 906-909. Available at: http://euralex.org/wp-content/themes/euralex/proceedings/Euralex 2016/euralex_2016_100_p906.pdf).

Wiegand, H.E. (2012). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung. Mit Berücksichtigung anglistischer, nordistischer, romanistischer, slavistischer und weiterer metalexikographischer Forschungen.* Berlin, Boston: De Gruyter Mouton.