Towards a Global Lexicographic Infrastructure

Simon Krek[1], Thierry Declerck[2,3], John McCrae[4], Tanja Wissik[3]
[1]Jožef Stefan Institute, Slovenia
[2]DFKI GmbH, Multilinguality and Language Technology, Germany
[3]Austrian Centre for Digital Humanities at the Austrian Academy of Sciences, Austria
[4]Insight Centre for Data Analytics at the National University of Ireland Galway, Ireland
[1]simon.krek@guest.arnes.si
[2]declerck@dfki.de
[3]Tanja.Wissik@oeaw.ac.at
[4]john.mccrae@insight-centre.org

Abstract

In this paper we briefly describe the European project ELEXIS (European Lexicographic Infrastructure). ELEXIS aims to integrate, extend and harmonise national and regional efforts in the field of lexicography, both modern and historical, with the goal of creating a sustainable infrastructure which will enable efficient access to high quality lexical data in the digital age, and bridge the gap between more advanced and lesser-supported lexicographic resources. For this, ELEXIS makes use of or establish common standards and solutions for the development of lexicographic resources and develop strategies and tools for extracting, structuring and linking lexicographic resources. This paper is a kind of summary of a more technical description of ELEXIS included in the proceedings of the Globalex 2018 workshop.

Keywords: eLexicography, Standards, Infrastructure

Résumé

Dans cet article, nous décrivons brièvement le projet européen ELEXIS (infrastructure lexicographique européenne). ELEXIS a pour objectif d'intégrer, d'étendre et d'harmoniser les efforts nationaux et régionaux dans le domaine de la lexicographie, à la fois moderne et historique. Le but est de créer une infrastructure durable qui permettra un accès efficace à des données lexicales de haute qualité à l'ère numérique et de combler le fossé entre les ressources lexicographiques les plus avancées et celles beaucoup moins développées. Pour ce faire, ELEXIS utilise ou établit des normes et des solutions communes pour le développement de ressources lexicographiques et élabore des stratégies et des outils pour extraire, structurer et relier les ressources lexicographiques. Cet article est une sorte de résumé d'une presentation plus technique d' ELEXIS, qui est include dans les actes du workshop Globalex 2018.

1. Introduction

The field of lexicography has a long tradition of proposing as accurate as possible descriptions of languages. As stated in (Køhler Simonsen, 2017): "Lexicography is a four thousand-year-old discipline and dictionaries have been an integral part of commerce and human cultural history for centuries".

Since the 1980s, lexicographers have started to utilize computers and to apply computational methods. Online dictionaries are no longer only a reference work but are also seen as platforms for supporting advanced search facilities. This emerging field of e-lexicography, nevertheless, is still not clearly shaped, and methods and workflows not yet fully agreed on. Michael Rundell (2015) for example describes the current situation of e-lexicography as being in a transitional phase. A quotation of Robert Lew stating that "It seems that the web community, while enthusiastically embracing the novelty of online collaboration, propagates the traditional model of lexicographic description"[1].

In recent years, however, new developments have emerged in the field of e-lexicography, like the eLex conference se-ries[2], which started in 2009, the Globalex initiative[3], which was established at eLex 2015 or the European Network of e-Lexicography (ENeL) COST action[4], which in 2013 brought together for the first time a large number of lexicographers to discuss issues related to the emergence of new technologies. ENeL was set up to improve the access for the general public to scholarly dictionaries and make them more widely known to a larger audience. In the context of this network, a clear need emerged for a broader and more systematic exchange of expertise, for the establishment of common standards and solutions for the development and integration of lexicographical resources, and for broadening the scope of application of these high quality resources to a larger community, including the Semantic Web, Artificial Intelligence, Natural Language Processing and Digital Humanities. This is where ELEXIS comes into play.

2. ELEXIS

ELEXIS (European Lexicographic Infrastructure) is fostering cooperation and information exchange among lexicographical research communities. The infrastructure is a granted project under the H2020-INFRAIA-2016-2017 call, with the topic "Integrating Activities for Starting Communities" and started in February 2018[5]. ELEXIS is building on infrastructures defined in other projects and initiatives, especially CLARIN[6] and DARIAH[7], which allow language or Digital Humanities resources (both tools and data) to be shared.

A key goal of the ELEXIS project is thus to enable stakeholders to link their existing lexicographic resources, either as dictionaries or as standalone lexical descriptions encoded, and so to create a huge multilingual registry, a kind of "Matrix Dictionary" that connects lexicographic resources across common concepts.

2.1 A Matrix Dictionary for ELEXIS

A key goal of ELEXIS is the creation of a "Matrix Dictionary", that is formed of links created between lexicographic resources in different languages, domains and forms, independently if the language considered is high- or under-resourced. With this, ELEXIS is creating a universal repository of linked senses, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in all types of existing lexicographic resources, monolingual, multilingual, modern, historical, etc.

In order to reach this goal, ELEXIS makes uses of strategies, tools and standards for extracting, structuring and linking the high-quality semantic data from lexicographic resources and make them available to the Linked (Open) Data[8] family.

Those processes are necessary, as current lexicographic resources, both modern and historical, have different levels of structure and are not equally suitable for applications in advanced NLP technologies, like Information Retrieval or Machine Translation, for which they should be disclosed to or from which they could benefit. The project works also on interlinking lexical content with other structured or unstructured data – corpora, multimodal resources, etc. – on any level of lexicographic description: semantic, syntactic, collocational, phraseological, etymological, translation equivalents, examples of usage, etc.

[1] Taken from taken from (Lew, 2014).

[2] See https://elex.link/

[3] See https://globalex.link/

[4] See https://www.elexicography.eu/

[5] See http://www.elex.is/

[6] See https://www.clarin.eu/.

[7] See https://www.dariah.eu/.

[8] See https://www.lod-cloud.net/ and also https://linguistic-lod.org/ for the subset of the LOD dealing with linguistic data.

ELEXIS conversion and alignment tools will provide users of the infrastructure with the possibility to harmonise and convert their lexicographic resources to a uniform data format that allows their seamless integration in Linked Open Data or in other repositories.

2.2 The virtuous Cycle of e-Lexicography in ELEXIS

ELEXIS implements a cyclic approach to the building and linking of lexicographic resources. We use the term "virtuous circle" for this, as it characterizes an integrative approach to a spiralling development of lexicographic data on the basis of a cross-disciplinary exchange of knowledge and the incremental contributions of the different methods and technologies to be involved. Figure 1 is given a graphical representation of this cyclic development.

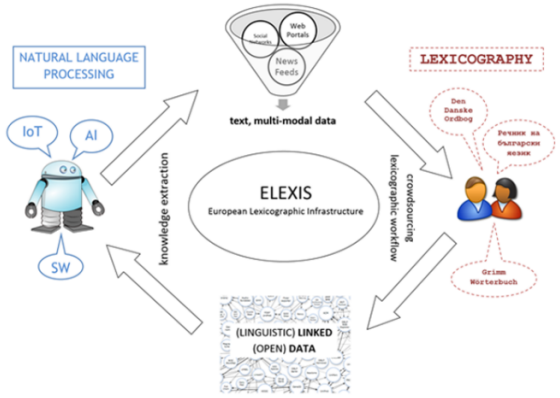

Figure 1: The virtuous cycle of e-lexicography

The existence of common data models and standards that are produced bottom-up from within the lexicographic community fostered by ELEXIS is a necessary condition for the successful development of the whole platform. Standards are developed and tested during the project on the data provided by the lexicographic partners and implemented in the newly developed service.

3. Relevance for Less-Resourced Languages

ELEXIS supports novel lexicography by providing lexicographers with tools and methods that help them create new resources. Using machine learning, data mining and information extraction techniques proto-dictionary content will be produced in an automated way. The automatically extracted data can then be used as a starting point for further processing either in a more traditional lexicographic workflow or through crowdsourcing platforms, making it easier to create new resources. This novel approach can be applied to any language for which there is data available on the web. This is particularly important for under-resourced languages with out-dated or non-existent language descriptions, enabling researchers and the general public to learn about semantic, grammatical or other aspects of lexicographic description benefiting from the technology derived from language communities with advanced lexicographic descriptions.

4. Summary

ELEXIS is aiming at the following points:

- 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;
- establish common standards and solutions for the development of lexicographic resources;
- develop strategies, tools and standards for extracting, structuring and linking of lexicographic resources;
- enable access to standards, methods, lexicographic data and tools for scientific communities, industries and other stakeholders;
- promote an open access culture in lexicography, in line with the European Commission Recommendation on access to and preservation of scientific information.

ELEXIS is based on the conviction that lowering the barrier for retrieving and analysing multilingual lexicographic data across Europe – and beyond cannot be accomplished in the long term without lowering the barrier for providing lexicographic data to research infrastructures.

5. Acknowledgements

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

6. Bibliographical References

Kilgarriff, A. (2000). Business models for dictionaries and nlp. International Journal of Lexicography, 13(2):107-118.

Køhler Simonsen, H. (2017). Lexicography: What is thebusiness model? In Iztok Kosem, et al., editors, Electronic Lexicography in the 21st Century, pages 395–415. Lexica Computing CZ s.r.o.

Lew, R. (2014). User-generated content (ugc) in online English dictionaries. OPAL - Online publizierte Arbeiten zur Linguistik, 4:8–16.

Rundell, M. (2015). From print to digital: Implications for dictionary policy and lexicographic conventions. Lexikos, 25(1).