# European Lexicographic Infrastructure (ELEXIS)

***Simon Krek*[1], *Iztok Kosem*[1], *John P. McCrae*[2], *Roberto Navigli*[5], *Bolette S. Pedersen*[6], *Carole Tiberius*[4], *Tanja Wissik*[3]**

[1]*Jožef Stefan Institute,* [2]*Insight Centre for Data Analytics, National University of Ireland Galway,* [3]*Austrian Academy of Sciences,*[4]*Dutch Language Institute,* [5]*Sapienza University of Rome,* [6]*University of Copenhagen*

*E-mail: simon.krek@ijs.si, john@mccr.ae, iztok.kosem@ijs.si, tanja.wissik@oeaw.ac.at, carole.tiberius@ivdnt.org, navigli@di.uniroma1.it, bspedersen@hum.ku.dk*

## Abstract

In the paper we describe a new EU infrastructure project dedicated to lexicography. The project is part of the Horizon 2020 program, with a duration of four years (2018-2022). The result of the project will be an infrastructure which will (1) enable efficient access to high quality lexicographic data, and (2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources. One of the main issues addressed by the project is the fact that current lexicographic resources have different levels of (incompatible) structuring, and are not equally suitable for application in in Natural Language Processing and other fields. The project will therefore develop strategies, tools and standards for extracting, structuring and linking lexicographic resources to enable their inclusion in Linked Open Data and the Semantic Web, as well as their use in the context of digital humanities.

**Keywords:** lexicography, research infrastructure, natural language processing, computational linguistics, semantic web, artificial intelligence, linked open data, digital humanities

## 1 Introduction

Reliable and accurate information on word meaning and usage is important in the information-driven society of the 21st century. In most European countries, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. Although confronted with similar problems relating to technologies for producing and making these resources available, cooperation on a larger European scale has long been limited. Consequently, the lexicographic landscape in Europe is rather heterogeneous. Firstly, it is characterized by stand-alone lexicographic resources, which are typically encoded in incompatible data structures due to the isolation of efforts, prohibiting reuse of this valuable data in other fields. Secondly, there is a significant variation in the level of expertise and resources available to lexicographers across Europe. Both issues contribute to the fact that the data from these resources is lost for extensive, interoperable and generally accessible computer use.

On the other hand, the language technology community, for their part, have created an overwhelming number of different types of lexical resources over the last thirty years, which are used for natural language processing tasks. These include corpora, lexicons, glossaries (used in machine translation), machine-readable dictionaries, lexical databases, and many others. One of the crucial issues addressed by ELEXIS is the fact that in the past the impressive results of the language technology community have rarely found their way into the practical work of creating lexicographic resources. This can be largely attributed to the lack of a common platform for building, sharing and exploiting knowledge and expertise between computational linguistics and lexicography, which is one of the goals of the proposed infrastructure.

In 2013, the European lexicographic community was brought together in the European Network of e-Lexicography (ENeL) COST action.[1] This initiative 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 has 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, NLP and digital humanities. At the end of the COST action, the initiative had been successfully transformed into a H2020 infrastructure project – European Lexicographic Infrastructure (ELEXIS).[2]

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.

## 2 The Consortium

The consortium is composed of content-holding institutions and researchers with complementary backgrounds in terms of lexicography, digital humanities, standardization, language technology, the Semantic Web and artificial intelligence, and it cooperates strongly with the existing CLARIN and DARIAH[3] infrastructures. ELEXIS project partners are:

1. The Jožef Stefan Institute, Slovenia (leading partner)
2. Lexical Computing, Czech Republic
3. Dutch Language Institute, Netherlands
4. Sapienza University of Rome, Italy
5. National University of Ireland, Galway, Ireland
6. Austrian Academy of Sciences, Austrian Centre for Digital Humanities, Austria
7. Belgrade Center for Digital Humanities, Serbia
8. Hungarian Academy of Sciences, Research Institute for Linguistics, Hungary
9. Institute for Bulgarian Language, Prof. Lyubomir Andreychin, Bulgaria
10. Universidade Nova de Lisboa, Faculty of Social Sciences and Humanities, Portugal
11. K Dictionaries, Israel
12. Institute for Computational Linguistics A. Zampolli, Italy
13. The Society for Danish Language and Literature, Denmark
14. University of Copenhagen, Centre for Language Technology, Denmark
15. Trier University, Centre for Computational Linguistics and Digital Humanities, Germany
16. Institute of the Estonian Language, Estonia
17. Spanish Royal Academy, Spain

1 www.elexicography.eu

2 http://www.elex.is/

3 Web sites: https://www.clarin.eu/, https://www.dariah.eu/.

Broadly speaking, work in the consortium focuses on three different types of activities: (1) joint research activities, (2) networking activities and (3) development of ELEXIS infrastructure through what is defined as "virtual access" and "trans-national access" activities. In the following sections we describe the three types of activities.

## 3 Research activities in ELEXIS

Research in ELEXIS is generally focused on two areas: lexicography and natural language processing. Lexicographic resources, both born-digital and retrodigitized, have different levels of structure and are not equally suitable for application in advanced NLP technologies. ELEXIS goal is thus to develop 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 family. In addition to linking lexicographic content, we also work 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. The ultimate goal is the creation of a universal registry/network of semantic relations used as a semantic intermediary language for global knowledge exchange, focused on difficult polysemous vocabulary (single-word and multi-word), modern and historical; the realization of a universal lexicographic metastructure, i.e. a matrix dictionary spanning across languages and time.

In order to motivate interoperability, we enable partners and other stakeholders to encode their data with common concepts from models such as the BabelNet (Navigli & Ponzetto 2012) and other Semantic Web models, such as DBpedia. To ensure that there is integration at even the most basic level, ELEXIS partners will define a minimal common data model capturing the basic concepts of a lexicographic resource such as entries (single-word, multi-word), senses, syntactic frames, etymologies, etc. and linguistic relationships such as synonymy/antonymy, translation, domain/region/register classification, relatedness, and so on that will be compatible with existing models used in the community, including TEI (Text Encoding Initiative), LMF (Lexical Markup Framework) and OntoLex-Lemon (McCrae et al. 2017), a model for modelling lexicon and machine-readable dictionaries, and linked to the Semantic Web and the Linked Data cloud.

Ultimately, research results in ELEXIS will be integrated in a platform that allows lexicographers to see the results coming from information extraction and corpus information, as well as crowdsourcing. This will enable lexicographers to make more detailed and consistent analyses of words in context. However, it is important to note that from the point of view of human-oriented lexicography, faced with abundance of data, the emphasis is on knowing what not to say and on how to say it efficiently. Therefore, methods and tools for visualization and presentation of lexicographic data are also extremely important, and will receive due attention. These research activities create a so-called virtuous cycle of cross-disciplinary exchange of knowledge and data, as shown in Figure 1.

The virtuous cycle of eLexicography includes experimental validation of the integrated LLOD data in Natural Language Processing tasks (lexicography for NLP) and the use of NLP for lexicography. As regards the former, the following tasks will be shown to benefit from the huge amount of multilingual information integrated in this project:

- **Multilingual Word Sense Disambiguation**, addressing the paucity of sense-annotated sentences. The ELEXIS lexicographic resources will be utilized to bootstrap large training datasets for WSD in dozens of languages.

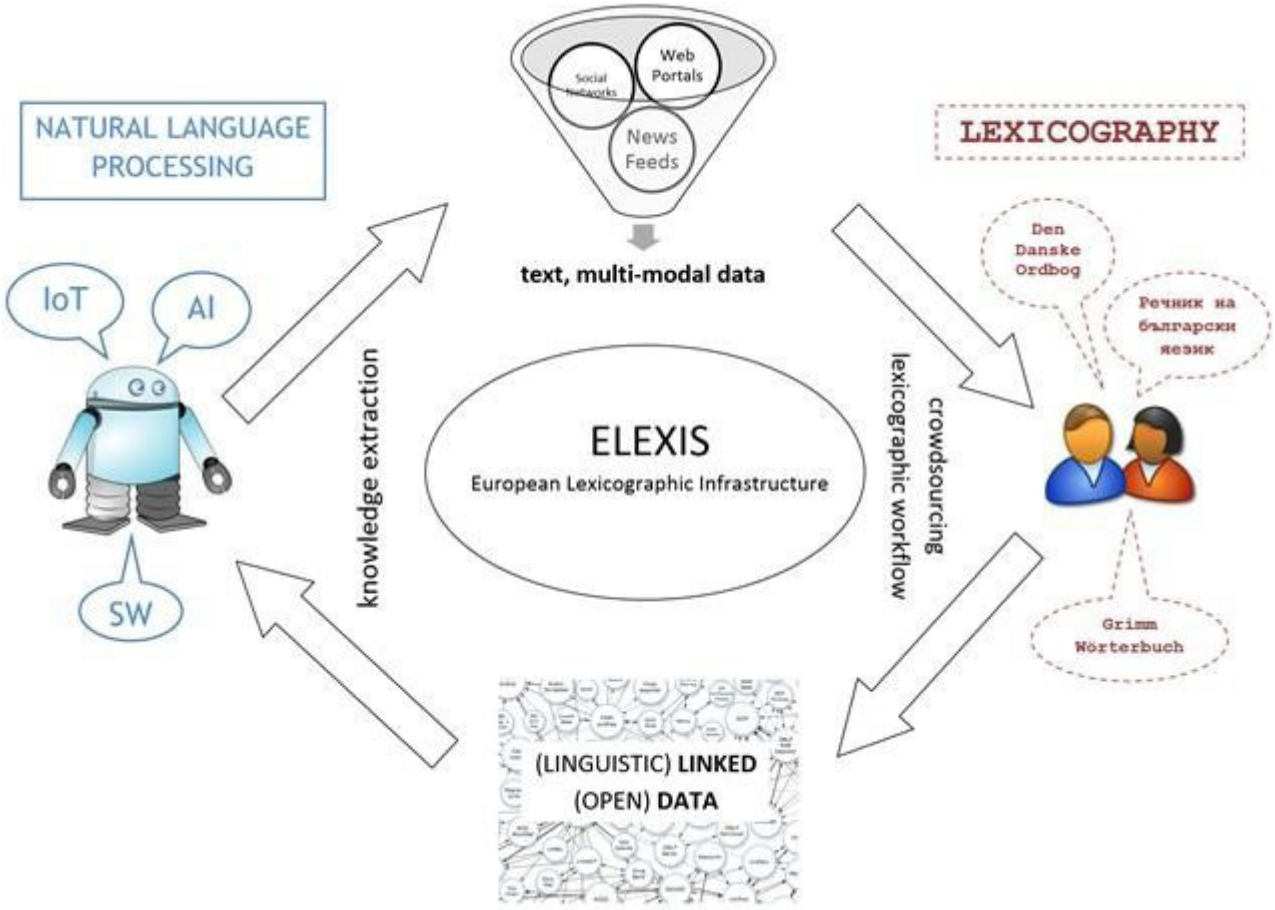


Figure 1: The virtuous cycle of eLexicography

- **Multilingual Semantic Parsing:** semantic parsing aims to map sentences to formal representations of their meaning. In ELEXIS we will develop innovative algorithms that exploit the huge multilingual network of interlinked lexical knowledge to perform multilingual semantic parsing.
- **Word sense clustering**, where the development of semi-automatic procedures to bring together subtle sense distinctions in clusters of meanings will be shown to improve the performance of tasks such as Word Sense Disambiguation;
- **Domain labelling of text**, where the aggregated information obtained from the lexicographic network of resources will be shown to improve automatic tagging of text with domain labels in arbitrary languages, thanks to developing innovative neural techniques.
- **Study of the diachronic distribution of senses**: the use of the most frequent sense in NLP is a solid baseline used in WSD and other tasks, but it is available only in the English language. We will develop novel techniques for aggregating the predominance information of senses a) from the multitude of resources and b) considering evolution over time, which will have an important impact on disambiguation and corpus analysis.

Advances in AI and NLP will, in turn, enable the development of improved tools for the production of structured proto-lexicographic data in an automated process, using machine learning, data mining and information extraction techniques resulting in a new type of lexicographic resource: a dictionary-on-the-fly. These new methods and tools will significantly shift the lexicographer's starting point and reduce the time-consuming parts of lexicographic work. In principle, having enough web or corpus data in a particular language will be a sufficient condition for a dictionary of that language

to be created, bridging the gap between lesser-resourced languages and those with advanced e-lexicographic experience. In addition to developing methods and tools for the automatic acquisition of lexicographic data, methods and tools for introducing crowdsourcing and gamification in the lexicographic process, and methods and tools to enrich lexicographic resources with multi-modal data will also be developed.

## 4 Networking Activities in ELEXIS

In order to reach the goals described in the introduction in Section 1, not only are research activities carried out, but networking activities are also needed, especially because ELEXIS will build and foster a community around the infrastructure and will support the exchange of knowledge. Therefore, special attention is given to networking activities that are reflected at different levels, as explained below.

### 4.1 Organization

The objective of ELEXIS is to foster cooperation and knowledge exchange among different research communities in lexicography in order to bridge the gap between less-resourced languages and those with advanced e-lexicographic experience, and one of the impacts of ELEXIS is defined as the emergence of a new type of lexicography that no longer views languages as isolated entities, but fully embraces the pan-European nature of those spoken in Europe. This ambition also extends to the global level. ELEXIS plans reflect this with an inclusive multi-layered organization that aims at engaging different user groups with various levels of intensity during the project, as shown in Figure 2.

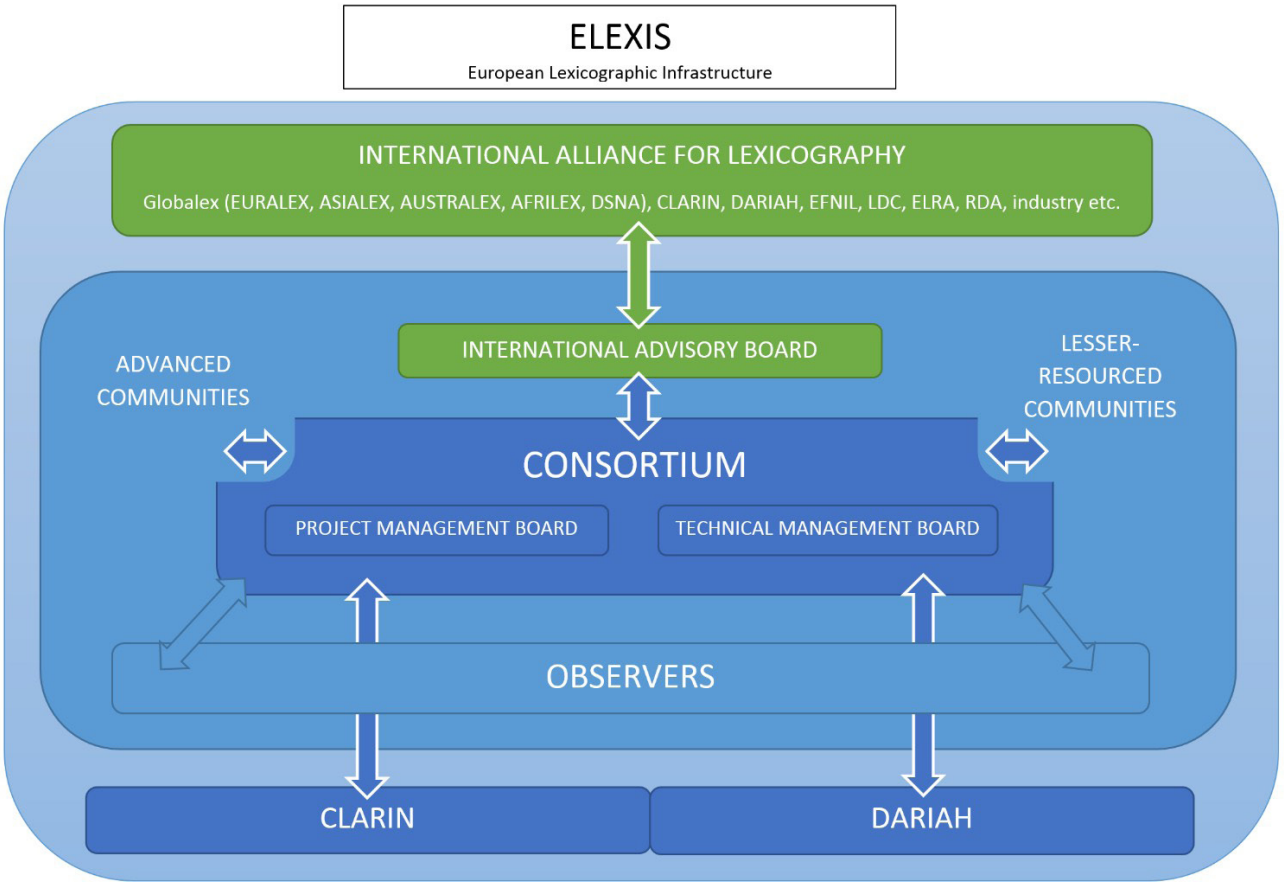


Figure 2: ELEXIS organization

### *4.1.1 The Consortium and Observers*

Elements of the structure consist of ELEXIS consortium partners as the core group responsible for the development of the infrastructure. Another organizational layer is observing institutions that will be directly included in outreach and dissemination activities through various channels. The central group of institutions that fall under the observer category are those producing quality lexicographic data and resources, filtered by the criteria of inclusion in the European Dictionary

Portal developed by the European Network of e-Lexicography COST action.[4] Typically but not exclusively, these institutions include (European) national language institutes, large dictionary publishers and other prominent producers of lexicographic data.

### *4.1.2 Advanced and Less-resourced Communities*

The broadest and also least defined user group consists of less-resourced communities and advanced communities working in lexicography. In practical terms, these include all researchers interested in lexicography who will gain access to newly developed data, tools and services through ELEXIS virtual access activities. The consortium itself reflects this division, as it includes partners working with less-resourced languages (Serbian, Slovenian, Irish, etc.) and from more advanced communities (English, German, Dutch, Italian, etc.). The infrastructure will provide networking activities (training, online training material, conferences, workshops, etc.) to enable researchers working with less-resourced languages to benefit from the results of the project, building on the expertise and data available in the more advanced communities.

### *4.1.3 International Alliance for Lexicography*

ELEXIS consortium partners will be advised by an International Advisory Board (IAB). Members of IAB include appointed representatives of five continental societies and associations for lexicography and top experts in both relevant fields, lexicography and NLP, from academia and industry. Through the International Advisory Board, and using its outreach and community building activities, ELEXIS will strive to form an International Alliance for Lexicography, which will include stakeholders from various fields. The alliance will be formed as a loose organization dedicated to the advancement of lexicography in the digital age, with the ELEXIS International Advisory Board serving as the initial governing body. It is expected that the alliance will be joined by the five continental lexicographic associations who form the Globalex initiative,[5] standardization bodies organized in EFNIL (European Federation of National Institutions for Language), data providers such as LDC (Linguistic Data Consortium), ELRA (European Language Resources Association), RDA (Research Data Alliance), and representatives of both language technology (Language Technology Industry Association – LT Innovate) as well as lexicography and language learning industries (e.g. Oxford University Press, MacMillan Publishers). The aim of the alliance is to consolidate the field of lexicography and enable its transition to the digital environment, bringing together all stakeholders interested in language description and semantic data.

### *4.1.4 CLARIN and DARIAH*

ELEXIS will also serve as a hub between CLARIN and DARIAH: dictionaries are essential language resources whose quality, reliability and coverage can be vastly improved by means of harmonizing formats and optimizing points of infrastructural access. At the same time, however, dictionaries are

4 Web page: http://www.dictionaryportal.eu/en/catalog/.

5 Web page: http://globalex.link/.

also objects of humanities research in their own right. Humanities scholars study dictionaries in terms of their cultural and ideological values, or their role in language standardization and nation-building, to name just a few different perspectives.

ELEXIS as a new infrastructure builds upon the existing tools and services of CLARIN and/or DARIAH with the goal of achieving something that neither infrastructure can at the moment provide on its own: a concerted pan-European effort aimed at combining and advancing the state-of-the-art in three distinct fields — lexicography, NLP and digital humanities. CLARIN already has a leading role in providing language resource repositories, linguistic annotation pipelines and federated search facilities, whereas DARIAH is a leader in facilitating long-term access to and use of arts and humanities research data. By creating a common platform for building, sharing and exploiting high-quality, multilingual lexical data, ELEXIS will aim to serve as a catalyst for closer cooperation between the two existing infrastructures. ELEXIS can succeed in this role because it has assembled a critical mass of eminent stakeholders from various disciplines who have both the technical and scholarly potential to: 1) help lexicographers build better dictionaries using the most advanced NLP techniques; 2) provide NLP researchers with high-quality lexicographic data to test and improve their algorithms on; and 3) aid humanities scholars in accessing social, historical and cultural data contained in legacy dictionaries in order to develop new procedures and tools for analyzing, visualizing and interpreting large sets of lexical data.

## 4.2 Dissemination and Community Building

Due to the nature of the ELEXIS infrastructure, various communities and types of users (professional, semi-professional and general public) will benefit on account of the scalable outcomes and services. We identified several major target groups, those who are undertaking lexicographic projects and those who are applying the high quality lexicographic data e.g. in the context of the Semantic Web, artificial intelligence, natural language processing and the digital humanities (Declerck et al. 2018). Next, we describe in more detail the different usage scenarios related to lexicographic projects.

### *4.2.1 Professional Large-scale Lexicography*

This group includes commercial and non-commercial entities undertaking large-scale lexicographic projects carried out by professional specialized teams from these areas:

- national language institutes (including consortium members)
- academia, universities, research institutions outside the ELEXIS consortium
- language standardization bodies and their umbrella organization EFNIL (European Federation of National Institutions for Language).
- industry (publishing houses, also software developers and language industry – in connection with large lexicographic projects).

### *4.2.2 Professional Small-scale Lexicography*

This group includes entities undertaking small-scale lexicographic projects carried out on a highly professional level either for research purposes or to address the needs of a small, well-defined community. The following groups may fall into this category:

- individual researchers (from the field of lexicography, also language studies, translation studies or the sister field of digital humanities, as well as natural language processing in connection with lexicography)
- trainers and students (interested in the educational aspects of the ELEXIS projects, such as learning material, training events)

- professionals and practitioners (language professionals, translators, proofreaders and others who use or produce linguistic resources in their daily professional life)
- freelance terminologists.

#### *4.2.3 Spontaneous and Small-scale Lexicography*

This group includes an enormous number of small projects often carried out without expertise in lexicography to address very specific needs of highly-specialized or very small professional or general public communities. A typical example would be a highly specialized domain-specific glossary. The following groups may fall into these categories:

- professional organizations, associations and authorities, non-profit organizations
- general public

Each group will be targeted with tailor-made messages that address the needs of the community to maximize the impact of any such activity.

### 4.3 Trans-national Access

During the lifetime of the project ELEXIS will organize trans-national calls enabling researchers (a) to work with data with restricted access at host institutions; and (b) to gain knowledge and expertise in close contact with lexicographers and experts in NLP and artificial intelligence. One of the reasons for the limited accessibility of lexicographic data outside institutions which are the creators and copyright holders of such data is the effort needed for their compilation, which necessitates tighter control over the access and availability of raw data. Trans-national activities represent one of the mechanisms of ELEXIS to enable access to such content for researchers from other institutions or countries. However, the results of research conducted in trans-national activities will be available under open access licenses according to the rules of the call enabling the international community to familiarize itself with previously inaccessible resources. In total, ELEXIS will give access to eleven European infrastructures/lexicographical milieus where researchers/lexicographers within the EU member states or associated countries are invited to apply for free-of-charge access via grant visits. During the visits, the hosting institutions will provide support in terms of both lexicographical and technical expertise.

## 5 ELEXIS Infrastructure

ELEXIS "virtual access" infrastructure – providing online access to data, tools and services – will consist of three sub-infrastructures: LEX1, LEX2, LEX3.

### 5.1 LEX 1

The first part of the infrastructure is dedicated to automatic segmentation and structuring of content for dictionaries that are currently produced in digital environments, but are typically encoded in their own custom data format. Conversion and alignment tools provide users of the infrastructure with the possibility to harmonies and convert their lexicographic resources into a uniform data format that allows their integration in Linked Open Data. Standards will be developed and tested during the project on the data provided by the lexicographic partners and implemented in the newly-developed service.

To provide conceptual interoperability, services enabling the linking of lexicographic resources will be developed and made available in the linking tools segment of the platform. This will provide the possibility to link lexical entries, senses and fundamental concepts in different lexical resources,

using a semi-automatic approach. BabelNet, as an existing multilingual resource to provide cross-lingual linking, is exploited for this purpose. Extensive linking of existing lexicographic resources by pivoting through BabelNet will enable the creation of what we call ELEXIS matrix dictionary – 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.

## 5.2 LEX 2

Based on the contribution of lexicographic data, a new infrastructure is being developed that includes word sense disambiguation and entity linking tools dedicated to semantic processing of corpus data. These tools will have an important impact on disambiguation and corpus analysis, and will open up the possibility to create lexicographic data from corpora in a fully automated process. This is included in the dictionary-on-the-fly segment of the platform. The service will be able to produce a proto-dictionary with sense distribution, extracted definitions, collocations, multi-word expressions, (good dictionary) examples, translation equivalents and data in other modalities.

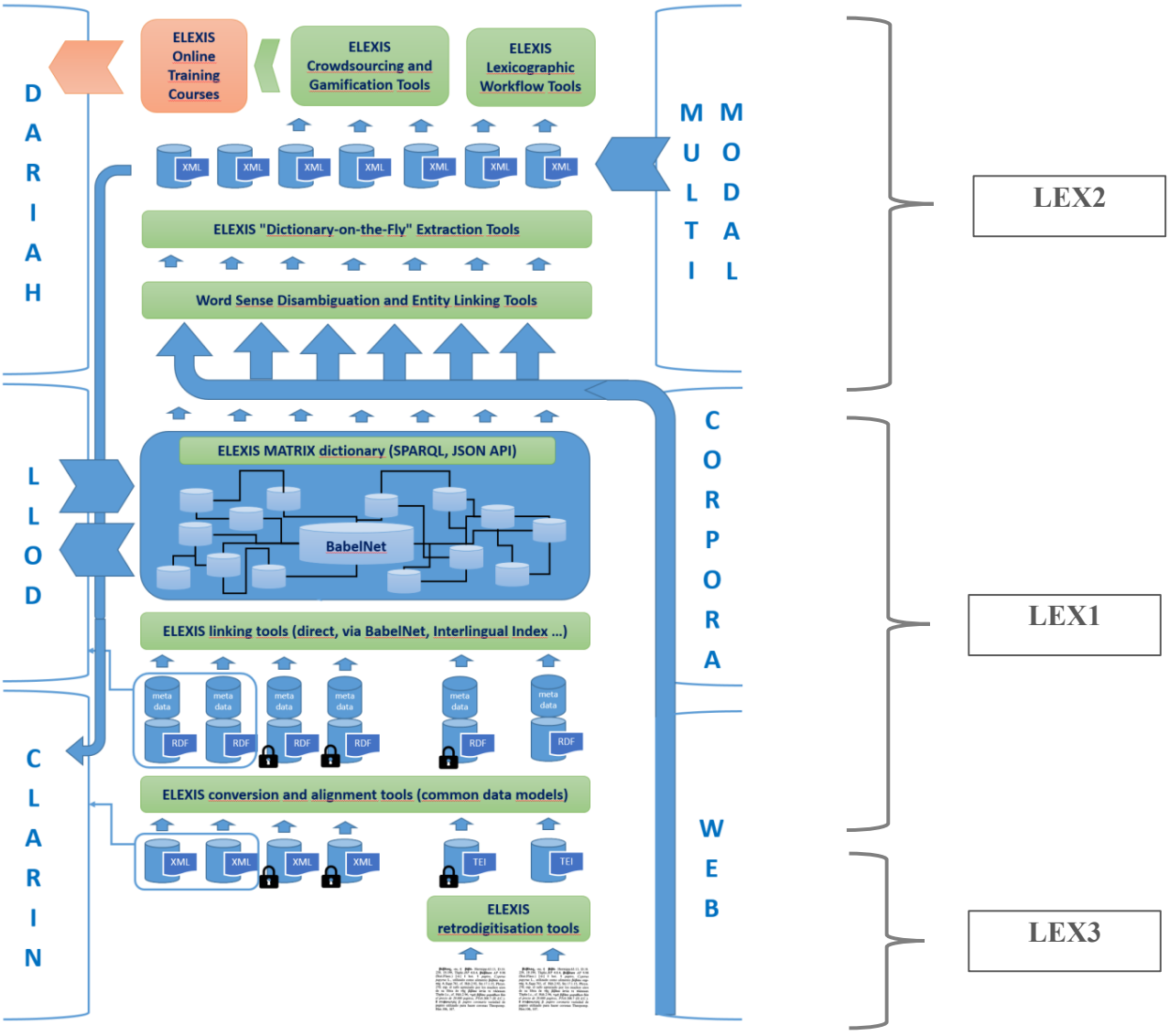


Figure 3: Tools and services: green, (lexicographic) data: blue, online training and education: brown.

To enable online lexicographic work on both existing and new (extracted) lexicographic data, two complementary sets of tools are provided: lexicographic workflow tools and crowdsourcing and gamification tools. The first include an open source online dictionary writing system, with the aim to provide the central dictionary writing platform which also includes new possibilities of online collaboration. The second provide tools for new techniques of dictionary creation, such as explicit or implicit crowdsourcing (gamification).

### 5.3 LEX 3

The third set of services is dedicated to retrodigitized dictionaries in the part of the platform that includes (1) tools for automatic segmentation and structuring of content in retro-digitized dictionaries, and (2) an online publication tool for retrodigitized dictionaries which also offers interfaces for the analysis and profiling of the underlying lexical data.

## 6 Conclusion

As a new infrastructure, ELEXIS brings together research communities and consortium partners working in different fields, in order to support the community working in the emerging field of e-lexicography. In particular, ELEXIS builds on the existing expertise and knowledge of partners in the fields of lexicography, computational linguistics and artificial intelligence in an interdisciplinary effort to make existing lexicographic resources available on a significantly higher level compared to their availability as stand-alone resources, which is the current state of affairs.

To support the lexicographic process and contribute to lexicography-oriented language description, ELEXIS will:

- develop methods and tools for the automatic processing and extraction of data from corpora and other (multimodal) resources for lexicographic purposes;
- develop methods and tools for the inclusion of extracted data into interlinked (open) lexicographic data;
- develop methods, guidelines and tools enabling the use of crowdsourcing and citizen science in the lexicographic process;
- elaborate on the guidelines and solutions for handling copyright and authorship protection to enable inclusion of extracted data into the lexicographic workflow.

To support the natural language processing community, several steps are needed to make existing lexicographic resources globally available. Therefore, ELEXIS will:

- develop methods, guidelines and tools for harmonization of dictionary formats, building on the existing standards within the lexicographic and NLP community;
- develop methods and tools for automatic segmentation and identification of dictionary structure, enabling interlinking of dictionary content;
- develop methods and tools for interlinking, maintenance, reuse, sharing and distribution of existing lexicographic resources;
- define evaluation and validation protocols and procedures (lexicographic data seal of compliance);
- elaborate on the guidelines and solutions for handling copyright and authorship protection to enable open access to lexicographic data in the LOD framework.

## References

Declerck, T. et al (2018)_....

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pp. 587-597.

Pilehvar, M. T., Navigli, R. (2014). A Robust Approach to Aligning Heterogeneous Lexical Resources. *Proceedings of ACL 2014*, pp. 468-478.

Navigli R., Ponzetto, S. (2012): BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, Elsevier, pp. 217-250.

## Acknowledgements

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