ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities

Bolette S. Pedersen[1], John McCrae[2], Carole Tiberius[3], Simon Krek[4]
University of Copenhagen, Denmark[1] , National University of Ireland Galway, Ireland[2], Dutch Language Institute, The Netherlands[3], Jožef Stefan Institute, Slovenia[4]
bspedersen@hum.ku.dk[1], john@mccr.ae[2], Carole.Tiberius@ivdnt.org[3], simon.krek@guest.arnes.si[4]

Abstract

The paper describes objectives, concept and methodology for ELEXIS, a European infrastructure fostering cooperation and information exchange among lexicographical research communities. The infrastructure is a newly granted project under the Horizon 2020 INFRAIA call, with the topic Integrating Activities for Starting Communities. The project is planned to start in January 2018.

1. Background

Reliable and accurate information on word meaning and usage is of crucial importance in today's information society. The most consolidated and refined knowledge on word meanings can traditionally be found in dictionaries – monolingual, bilingual or multilingual.

Dictionaries are not only vast, systematic inventories of information on words, they are also important as cultural and historical artefacts. In each and every European country, 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 currently rather heterogeneous. On the one hand, it is characterised by stand-alone lexicographic resources, which are typically encoded in incompatible data structures due to the isolation of efforts, complicating reuse of this valuable data in other fields, such as natural language processing, linked open data and the Semantic Web, as well as in the context of digital humanities. On the other hand, there is a significant variation in the level of expertise and resources available to lexicographers across Europe. This forms a major obstacle to more ambitious, innovative, transnational, data-driven approaches to dictionaries, both as tools and objects of research.

In 2013, the European lexicographic community was brought together for the first time in the European Network of e-Lexicography (ENeL) COST action (www.elexicography.eu). 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. This networking initiative, which ended in October 2017, started with 34 members from 20 countries but grew to 285 members from 31 countries. 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, NLP and digital humanities.

As a synthesis of the ENeL efforts, a consortium was established in 2016 and in August 2017 the proposal for an infrastructure under the name ELEXIS was selected for funding in the Horizon 2020 INFRAIA call with the topic Integrating Activities for Starting Communities. The project is planned to start in January 2018.

In the following sections we will outline the objectives, concept and methodology of the infrastructure, and finally we will sketch out some foreseen wordnet related research tasks in the project concerning sense clustering and multilingual linking.

2. Objectives

The main objectives of ELEXIS can be summarized as follows:

- to foster *cooperation* and *knowledge exchange* between different research communities in lexi-

cography in order to reduce the gap between lesser-resourced languages and those with advanced e-lexicographic experience; and
- to work with strategies, tools and standards for *extracting*, *structuring* and *linking* of lexicographic resources;
- to facilitate *the access* to standards, methods, lexicographic data and tools for scientific communities, industries and other stakeholders;
- to encourage to 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 cannot be accomplished in the long term without lowering the barrier for providing lexicographic data to research infrastructures. As a result, the following impacts are pursued:

- efficient (open) access to high quality lexicographic data for researchers, institutions and stakeholders from different fields;
- a common platform for building, sharing and exploiting knowledge and expertise between lexicography and computational linguistics which will facilitate cross-disciplinary fertilisation and a wider sharing of information, knowledge and technologies across and within these fields. The platform will thus aim at bridging the gap between lesser-resourced languages and those with advanced e-lexicographic and/or computational linguistic experience;
- the creation of a scalable, multilingual and multifunctional, language resource. By integrating and linking lexical content and interlinking it with other structured or unstructured data - corpora, multimodal resources, etc. - on any level of lexicographic description, the project will strive towards creating a multilingual and multifunctional language resource incrementally enriching the available information;
- the inter- and multidisciplinary nature of lexical data will help researchers ask new questions and pursue new avenues of research.

3. ELEXIS Participants

The ELEXIS consortium includes the following 17 participants:

1. "Jožef Stefan" Institute, Slovenia
2. Lexical Computing, Czech Republic
3. Dutch Language Institute, The Netherlands
4. Sapienza University of Rome, Italy
5. National University of Ireland, Galway, Ireland
6. Austrian Academy of Sciences, 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. Faculty of Social Sciences and Humanities, Universidade Nova de Lisboa, Portugal
11. K Dictionaries, Israel
12. Consiglio Nazionale delle Ricerche - Istituto di Linguistica Computazionale "A. Zampolli", Italy
13. The Society for Danish Language and Literature, Denmark
14. University of Copenhagen, Denmark
15. Trier University, Center for Digital Humanities, Germany
16. Institute of Estonian Language, Estonia, and
17. Real Academia Española, Spain

4. Concept and Methodology

ELEXIS will build 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 to a certain degree the current state of affairs.

These resources are in fact results of long-term projects in which literally thousands of person years were and continue to be dedicated to their compilation in national and regional projects, and in most cases they represent the most consolidated and refined knowledge on word meanings in individual languages. A tremendous effort is needed for their compilation, and this implies the necessity to control the contents in order to ensure both the continuation of consistent language description and maximum quality of the results. Furthermore, and resulting from current isolation of efforts, these resources are typically encoded in incompatible data structures. Both issues con-

tribute to the fact that the data from these resources is currently not fully accessible for extensive, interoperable computer use.

On the other hand, the language technology (LT) community, for their part, 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 important issues that will be addressed by ELEXIS is the fact that the impressive results of the LT community have only to a limited degree found their way into the practical work of creating lexicographic resources in the past. 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 ELEXIS infrastructure.

4.1 Supporting lexicographic process and language description

To support the lexicographic process and to contribute to lexicographically-oriented language description ELEXIS will work towards:

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

4.2 Supporting natural language processing

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

- develop methods, guidelines and tools for harmonisation 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 LOD framework.

Therefore, in contrast with previous more NLP-oriented efforts spanning from computational lexicographical projects like EAGLES (Calzolari et al. 2002), PAROLE/SIMPLE (Lenci et al. 2000) to a current infrastructure on language resources and technology like CLARIN, ELEXIS will develop methods and tools to produce collections of structured protolexicographic data in an automated process, using machine learning, data mining and information extraction techniques, where the extracted data can be used as a starting point for further processing either in the traditional lexicographic process or through crowdsourcing platforms. In this context, focus will be on defining interoperability standards and data services in close cooperation with the existing CLARIN and DARIAH infrastructures.

4.3 Methodology

Lexicography as a field has a long tradition of refining semantic description of individual languages in comprehensive monolingual dictionaries, or performing detailed contrastive analysis between two or more languages in bilingual and multilingual dictionaries. However, these resources are currently not used to a sufficient degree within existing and emerging language technologies. They are almost completely absent in linked (open) data clouds and Semantic Web technologies, and are to some degree "digitally invisible".

In the last decade the new field of e-lexicography emerged, which can be seen in initiatives such as the ENeL COST action (http://www.elexicography.eu/), the eLex conference series (https://elex.link/), or the Globalex workshop at LREC 2016 (http://ailab.ijs.si/globalex/). Globalex is the first

initiative which includes all continental lexicographic associations: EURALEX, ASIALEX, AFRILEX, AUSTRALEX and the Dictionary Society of North America. The field of e-lexicography is dedicated to creating digitally-born dictionaries defined as lexical resources intended for human users but intentionally moving away from the paper medium and exploring the almost infinite possibilities of the new digital environment, with a view to take human-oriented lexical description to entirely different levels. In this context, machine learning, data mining and other computational techniques are starting to find their way into lexicography. Combining both traditional lexicographic knowledge and expertise with computational linguistics, while engaging also wider language communities in the process, creates huge potential for the development of the field.

4.4 Lexicography and »semantic bottleneck«

Lexicographic resources contain quality information about general vocabulary and more difficult types of language phenomena such as highly polysemous words or semantically opaque multi-word expressions (idioms, phraseology), which are rather inconsistently covered in LT-oriented resources. These phenomena represent a bottleneck in achieving precision and computational efficiency of NLP applications. This can be seen also from efforts such as PARSEME COST action (http://typo.uni-konstanz.de/parseme/) which was devoted to the role of multi-word expressions in parsing. Word sense disambiguation as part of content analytics, text understanding and computer reasoning remains another complex task for computational processing of text, and is still largely unsolved, especially for languages other than English. Typically, resources such as Wikipedia, Wiktionary, wordnets or framenets are used for word sense disambiguation tasks, collected in the (L)L(O)D cloud (http://linked-data.org/, http://www.linguisticlod.org/). Knowledge bases and complementary applications such as BabelNet (http://babelnet.org/), Babelfy (http://babelfy.org/), Cyc (http://sw.opencyc.org/) or wikifiers (http://-www.wikifier.org) have been developed to enrich text processing with semantic information. ELEXIS proposes enriching the existing linked data clouds and knowledge bases with data available in existing and new lexicographic resources, which are currently not used for solving these tasks.

4.5 Standards in lexicography and NLP

There are several reasons for the negligible incorporation of lexicographic data in LT so far. The first is almost non-existent interoperability and use of common standards in lexicography. In past decades there were several important efforts to harmonise and standardise linguistic resources, including lexicographic resources. These include first initiatives such as EAGLES/ISLE, Multext(-East), PAROLE, SIMPLE, CONCEDE etc. in the 1990s. From these efforts, standards emerged such as Text Encoding Initiative (TEI - http://www.tei-c.org/), Lexical Markup Framework (LMF - http://www.lexicalmarkupframework.org/), and others, most of them under the umbrella of the Terminology and other language and content resources ISO/TC 37 standard.

The standardisation process was much more successful with resources directly dedicated to computer use, such as corpora, lexicons, lexical databases, wordnets, ontologies etc., but standards were less successful in case of lexicographic resources initially intended for human users.

4.6 Availability of lexicographic data

Although early digitisation projects involving lexicographic resources date back to the 1980s (Boguraev and Briscoe, 1989), or in case of English even the 1960s (Urdang, 1966), and even if the 1990s saw massive digitisation of existing dictionaries, including works like the Oxford English Dictionary, general (open) access to lexicographic data is extremely limited. The main reason for this is the massive effort necessary to compile such resources, either by national language institutions, or by commercial companies in the case of "commercial languages" with a sufficient number of speakers.

The effort needed consequently implies the necessity to control the contents, resulting in the need to resolve intellectual property right issues before this data can be included in open access infrastructures.

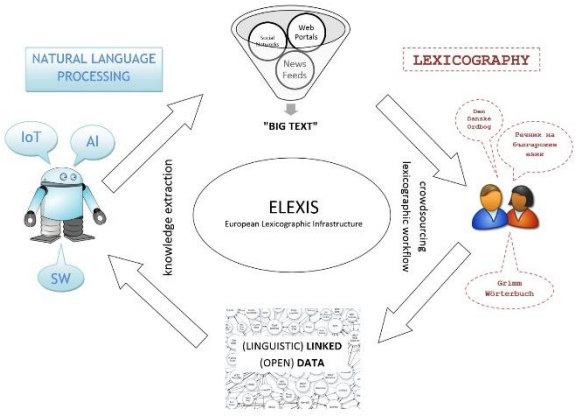
The ELEXIS infrastructure will dedicate serious efforts to handle IPR issues related to lexicographic data and enable their integration as linked data. In the last decade, initiatives promoting open access to the results of publicly funded projects (Open Research Data Pilot etc.) and the increasing wealth of (open) data available on the Web (Wikipedia, Wiktionary etc.), also instigated new trends within lexicography,

particularly the move towards e-lexicography. This new trend is not yet supported by an infrastructure where quality semantic data from dictionaries could be linked, shared, distributed and stored on a massive scale. Therefore, the objective of contributing quality semantic data in the digital age means that the proposed project will work towards enabling existing lexicographic resources to be included seamlessly into the Linked (Open) Data family (see Picture 1).

4.7 Virtuous cycle of e-lexicography

As was established in ENeL surveys, the results of the LT work are rarely used in lexicography, which is one of the important issues addressed by the ELEXIS infrastructure. This can be largely attributed also to the lack of an infrastructure enabling sharing knowledge and expertise between LT and lexicography. Ideally, the part of the virtuous cycle starting from NLP towards lexicography will produce proto-dictionary content in a completely automatic process with the use of machine learning, data mining and information extraction techniques focusing on massive amounts of data in various modalities available on the Web.

ELEXIS aims to develop methods and tools to produce such collections of structured data in an automated process where the data can be used as a starting point for further processing of the collected material either by traditional lexicographic process or through crowdsourcing platforms.



Picture 1: Virtuous cycle of e-lexicography

5 ELEXIS and Wordnets

5.1 Sense clustering and predominance information

Based on achievements from BabelNet (Navigli et al. 2012) and other works as found in Izquirdo et al. 2009, McCarthy et al. 2016, and Pedersen et al. forthcoming, ELEXIS will work on developing principled methods for sense clustering which are preferably semi-automatic. This also involves wordnet sense inventories which in many cases incorporate ontological typing but are on the other hand not organised in main- and sub-senses (in opposition to most dictionaries).

The project will include frequency and predominance information of senses in this work with the overall aim of improving word sense disambiguation and other NLP-related tasks. The intention is to work towards making existing lexical resources including wordnets more operational and practically useful in NLP by focusing on the organisation of the sense inventory.

Frequency and predominance information of senses is however not information which is directly accessible for all the involved languages at the current stage. Therefore, an initial task will be to develop methods to process these data for the less-resourced European languages.

5.2 ELEXIS and the WordNet Interlingual Index

The Global WordNet Association has proposed an Interlingual index of concepts (Bond et al., 2016), in which synsets from any wordnet can be identified with a single unique identifier, enabling interlingual linking of wordnets. It is clear that these goals correspond well with those of the ELEXIS project and it is expected that the benefits of these tools will be offered also to the wordnet community.

As a minimal step to enable this, the XML LMF format of the Global Wordnet Association[1] will be supported as a valid input and output format to the tools developed in the context of ELEXIS. Thus the linking tools that will establish cross-lingual similarity between concepts will be applicable to wordnets and thus this will be used to detect duplicate concepts between different wordnets and ameliorate the task of introducing new interlingual identifiers. Secondly it is hoped that the knowledge extraction components of the ELEXIS infrastructure will be integrated

[1] http://globalwordnet.github.io/schema

into the lexicographic procedures for new wordnet creation, and we intend to demonstrate this by integrating the crowd-sourced procedure used to create the Colloquial Wordnet (McCrae et al., 2017) within the ELEXIS infrastructure. In particular, this resource selects potential neologisms by NLP analysis of Twitter in order to detect terms that appear to be emerging in the English language.

Finally, it is expected that ELEXIS will encourage a closer interaction between the BabelNet project (Navigli & Ponzetti 2012) and other wordnets. In particular, BabelNet and the Interlingual Index will be linked so that users may access the data through either interface and new concepts in either resource can be integrated automatically. Furthermore, we will define metadata such that the licensing and sources for information can be clearly and unambiguously identified.

Acknowledgement

This work was supported by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

References

Boguraev B. and Briscoe T. (Eds.). (1989). *Computational Lexicography for Natural Language Processing*. Longman Publishing Group, White Plains, NY, USA.

Bond F., P. Vossen, J. McCrae, Ch. Fellbaum (2016). CILI: the Collaborative Interlingual Index, in: *Proceedings of the 8th Global WordNet Conference 2016* (GWC2016) in Bucharest, Romania, January 27-30.

Calzolari N., Zampolli A., Lenci. (2002). Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative. In: *CICLing 2002: Computational Linguistics and Intelligent Text Processing* pp 264-279

Izquirdo, R., A. Suárez, and G. Rigau. (2009). An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics pp 389-397*. The Association for Computational Linguistics.

Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, A. Zampolli. (2000). SIMPLE: A general framework for the development of multilingual Lexicons. International Journal of Lexicography, 13(4), 249-263

McCarthy, D., M. Apidianaki & K. Erk (2016). Word Sense Clustering and Clusterability. In: *Computational Linguistics*, Vol. 42, no. 2.

McCrae J.P., Wood I., Hicks A. (2017) The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In: Gracia J., Bond F., McCrae J., Buitelaar P., Chiarcos C., Hellmann S. (eds.) *Language, Data, and Knowledge*. LDK 2017. Lecture Notes in Computer Science, vol. 10318. Springer, Cham.

Navigli, R. and S. P. Ponzetto. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193: 217-250.

Pedersen, B.S., M. Agirrezabal, S. Nimb, S. Olsen, I. Rørmann (forthcoming). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In: *Proceedings of Global WordNet Conference 2018*, Singapore.

Urdang L. (1966). The Systems Designs and Devices Used to Process: The Random House Dictionary of the English Language. *Computers and the Humanities* 1 (2).