ELEXIS
European Lexicographic Infrastructure

Simon Krek
Jožef Stefan Institute / Centre for Language Resources and Technologies, University of Ljubljana, Slovenia

XVIII Euralex congress, 17th-21st July 2018
eLex 2011 (10-12 November 2011 – Bled, Slovenia)

Text mining is a challenge

- Automatic – detection, extraction, retrieval, selection, etc.
- Types of information:
  - definitions
  - collocations
  - examples
  - synonyms
  - multiword expressions
  - phraseology
  - etymology etc. etc.

world wide web
digitized present
digitized past
eLex 2011 – text mining (floating dictionary)

Jellyfish

Sinclair: Floating dictionary (2001)

- »A few years ago I felt that the time was ripe to plan a new kind of dictionary, one that would never exist on paper, but would be automatic or almost automatic in its self-updating.
- It would, so to speak, float on top of a corpus, rather like a jellyfish, its tendrils constantly sensing the state of the language.
- As well as reporting on the settled usage and meanings of the words and phrases of a language, like a normal dictionary does, the floating dictionary, when interrogated, dips into the corpus and checks this information, offering instances that match its criteria for the senses; also it explores further to see if there are any instances that conflict with the criteria, and may signify a development of a sense or the emergence of a new usage altogether.
- Within the limits of its powers, it organises this evidence as a comment on the existing dictionary entry.«
WordNet-LMF format

Text Encoding Initiative

- Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form.
- **9 Dictionaries:** This chapter defines a module for encoding lexical resources of all kinds, in particular human-oriented monolingual and multilingual dictionaries, glossaries, and similar documents.
- The elements described here may also be useful in the encoding of computational lexica and similar resources intended for use by language-processing software; they may also be used to provide a rich encoding for wordlists, lexica, glossaries, etc. included within other documents.
eLex 2011 – presentation

„Question answering“ mode

Expressing the need

• input type
  – question answering
  – conversational agents etc.

• input mode
  – mouse, keyboard (physical, touchscreen)
  – voice recognition (Apple, Android)
  – OCR (Abbyy)
  – gesture (Kinect) etc.
eLex 2011 – conclusions

... „and now for something completely different“

Conclusions

• A new interest in language technology is emerging, related to the requirements of the information society
• EU seems to be preparing a coordinated action to get back on track after the success of North American companies
• Is it possible to jump on the bandwagon, with a conceptual break with tradition in lexicography?
European Network of e-Lexicography (ENeL)
Kick-off meeting, 11 October 2013, Brussels

Working Groups

- **WG 1**: Integrated interface to European dictionary content
- **WG 2**: Retro-digitized dictionaries
- **WG 3**: Innovative e-dictionaries
- **WG 4**: Lexicography and lexicology from a pan-European perspective

Observations

- Prevalence of user-generated dictionaries/growing gap between scholarly dictionaries and the general public;
- Lack of common standards and solutions for retrodigitized dictionaries;
- Lack of a common research paradigm, common standards and solutions for e-lexicography;
- In dictionaries languages are treated as isolated entities.
European Network of e-Lexicography (ENeL)

Final meeting, 18 September 2017, Leiden

- 280 members
- 30 countries
- 7 meetings
- 4 workshops
- 37 STSMs
- 3 training schools
- 1 portal
- > 1 EU-funded project
ELEXIS FACT SHEET

• Call & Topic: INFRAIA-02-2017
  • Integrating Activities for **Starting Communities** (Publication: October 2015)
  • Model: two-stage (March 2016, March 2017), results: August 2017

• Start date: **1 February 2018**

• Duration: **48 months** (31 January 2022)

• Total cost: **4,999,967.50 €**

• Coordinator: Jožef Stefan Institute, Ljubljana, Slovenia

• Number of partners: 17 from 15 countries

• Web site: [www.elex.is](http://www.elex.is)
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

• Partners with:
  • lexicographic data and/or expertise
  • computational linguistics data and/or expertise
  • expertise in standardisation
  • digital humanities partners
  • technology partners
GOALS

• 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
  • (1) enable efficient access to high quality lexical data in the digital age, and
  • (2) bridge the gap between more advanced and lesser-resourced scholarly communities working on lexicographic resources.
EXPECTED IMPACTS

• Providing efficient **access to** quality lexicographic **data**
• Enabling massive **integration of** knowledge-based **resources**
  
  • Facilitating inclusion of innovative lexicography in **research** and **education**
  • Enabling the use of new technology and data in **industry**
  • Establishing **inter-infrastructure** synergies and optimisation

• **RESULT:** a **new** type of **lexicography** that no longer views languages as isolated entities
VIRTUOUS CIRCLE/CYCLE OF E-LEXICOGRAPHY

• “In the best of all possible worlds, computational enhancement and lexicographical upgrading would build upon each other in a virtuous circle that knew no bounds”.
  
  • Abstract: “NLP needs dictionaries, and dictionary-makers can use NLP to make better dictionaries, so there is great potential for synergy between the two activities. To date, there has been only very limited collaboration. This is substantially owing to dictionary publishers’ concerns regarding intellectual property. In this paper I explore the different interests of publishers and NLP researchers, and present a business model which pays heed to both.”

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
THREE TYPES OF ACTIVITIES

• Joint research activities
  • tools and methods for enabling linked lexicographic resources
  • tools and methods to support innovative e-lexicography

• Networking activities
  • documentation, guidelines, collections of best practices and multilingual data
  • workshops, seminars, and conferences
  • online training modules
  • data seal of compliance for lexicographic data
  • international forum

• Trans-national access or virtual access activities
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
What happens with your dictionary?

• (It gets (retro)digitised)

• You are using your own data model for your dictionary (Word, XML etc)
  • CONVERSION TO COMMON DATA MODEL (e.g. TEI, TEI-LEX-0)

• The human oriented common data model is not really NLP friendly
  • CONVERSION TO MACHINE READABLE FORMAT (e.g. Ontolex, Lemon)

• Now it can be linked with other dictionaries via shared concepts
  • LINKING DIRECTLY OR VIA BABELNET

• Shared concepts end up in MATRIX dictionary together with links to your original dictionary
  • MATRIX DICTIONARY, OPEN ACCESS RESOURCE WITH LINKED DATA
What happens with your corpus?

• It is processed on lower levels (POS-tagging, parsing etc.) and babelfied or wikified
  • WORD SENSE DISAMBIGUATION BASED ON BABELNET OR WIKIPEDIA
• now some methods can be applied to explore semantic behaviour of the vocabulary (also MWEs), to look for translation equivalents etc.
  • SEMANTIC ANALYTICS AND MULTILINGUAL SEMANTIC PARSING
• also, the corpus and the dictionary start functioning almost as one resource in push/pull model
  • ENRICHMENT OF LEXICOGRAPHIC RESOURCES
• the automatically extracted or enriched data can be manually curated in various dictionary writing systems, crowdsourcing platforms or in games
  • DWS, CROWDSOURCING TOOLS, GAMIFICATION
What is in the middle?

• 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.

• a universal lexicographic metastructure; a matrix dictionary spanning across languages and time
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
SKETCH ENGINE & Lexonomy (LCL)

• **Sketch Engine**: [https://www.sketchengine.eu/](https://www.sketchengine.eu/)
  • corpus query, corpus building and corpus management system allows users to build and work with 300+ text corpora in over 90 languages and 20 scripts
  • access funded through ELEXIS between **1 April 2018** and **1 April 2022**
  • access is provided to the **institution** and applies to non-commercial use
  • all **academic institutions** (universities, academies etc.) in EU can gain access
  • must have **SSO**, be member of **eduGAIN** and fill in the enrollment form

• **Lexonomy**: [https://www.lexonomy.eu/](https://www.lexonomy.eu/)
  • cloud-based dictionary writing and online dictionary publishing system
    (integration with the Sketch Engine: one-click dictionary)

• **Already available!**
TRANS-NATIONAL ACCESS

• ELEXIS will announce **visiting grants** with an average duration of two weeks for researchers to experiment with lexicographical data

• 5 calls for visiting grants will be launched during the project period amounting to an overall number of **30–40 grants**
  • the **calls** will include descriptions of the particular lexicographical resources, tools, and expertise that are made available
  • interested researchers/lexicographers should make an **application**, describing background, purpose, etc., which will be assessed by a committee
  • during the grant visits, the hosting institutions will provide **support** in terms of lexicographical and IT manpower expertise

• **1st call is out!** Deadline: 6 August 2018
HOSTING INSTITUTIONS

- **ELEXIS-SI**: Jožef Stefan Institute (JSI, Slovenia)
- **ELEXIS-NL**: Dutch Language Institute (INT, The Netherlands)
- **ELEXIS-EE**: Institute for Estonian Language (EKI, Estonia)
- **ELEXIS-DK**: Det Danske Sprog- og Litteraturselskab (DSL, Denmark) & University of Copenhagen (UCPH, Denmark)
- **ELEXIS-ES**: Real Academia Española (RAE, Spain)
- **ELEXIS-BG**: Institute of Bulgarian Language Lyubomir Andreychin (IBL, Bulgaria)
- **ELEXIS-HU**: Hungarian Academy of Sciences (RILMTA, Hungary)
- **ELEXIS-AT**: Austrian Academy of Sciences (OEAW, Austria)
- **ELEXIS-RS**: Belgrade Center for Digital Humanities (BCDH, Serbia)
- **ELEXIS-DE**: Trier Center for Digital Humanities (TCDH, Germany)
- **ELEXIS-IL**: K Dictionaries (KD, Israel)
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
OBSERVERS

• Rights
  • Inclusion in all ELEXIS activities (without direct funding)
  • Inclusion in communication channels
  • Access to (open) data and tools (for research purposes)
  • Active promotion of participation in trans-national access calls
  • Opportunity to participate in the ELEXIS post-project developments

• Obligations
  • Providing lexicographic data or offering expertise needed in ELEXIS
  • Providing feedback on lexicographic practices, workflows and tools

• Application form on the web page!
WHO CAN BE AN OBSERVER?

- Observer status can be granted to research institutions or other legal entities that are
  - involved in **lexicography** or
  - active in the fields connected with the goals defined in ELEXIS project – **natural language processing, artificial intelligence** etc.

- Institutions wishing to become ELEXIS observers can use this form to request observer status.
Inclusion in ELEXIS activities

• In the course of the project, observers will be offered different types of direct involvement in ELEXIS activities.

• Typically, such activities will include work with data, tools or services either provided by observers and enhanced or upgraded by ELEXIS infrastructure, or developed within ELEXIS for particular languages that are interesting for observers.

• Opportunities for such activities will be announced through internal communication channels for observers.
Inclusion in communication channels

• ELEXIS will develop various communication channels, both external (accessible to general public) and internal (dedicated to discussion about different ELEXIS activities).

• Observers will be informed about the possibilities and will be able to choose which types of information they would like to receive on a regular basis.
Access to (open) data and tools

• ELEXIS infrastructure will set up its own cloud infrastructure with lexicographic and other types of data, and tools developed by ELEXIS partners or as part of the ELEXIS infrastructure.

• Observers will be able to access parts of the cloud with data and tools that their owners will make available to the consortium partners and observers, together will open-access or open-source data and tools.
Active promotion of participation in trans-national access calls

• ELEXIS will organise several calls for grants to visit ELEXIS partners with lexicographic data and expertise.

• Observers will be notified about the calls offering the possibility to work on the data together with experts from other countries and lexicographic communities.

• Trans-national visit grants are especially suitable for observers that work on under-resourced languages.
Opportunity to participate in the ELEXIS post-project developments

• In the four years of the EU-funded ELEXIS project, the consortium aims to create a sustainable infrastructure that will serve lexicographic community also after the end of the project.

• One of the tasks in ELEXIS project is dedicated to finding a solution for long-term sustainability, together with the existing CLARIN and DARIAH infrastructures.

• Observers will be included in this task and will be able to participate actively in the discussion.
Obligations: Providing lexicographic data or offering expertise needed in ELEXIS

• Observers are expected to offer their contribution to ELEXIS project and infrastructure, either in the form of (a sample of) lexicographic/lexical data created at the observer institution, or by providing a type of expertise required by ELEXIS.

• Expertise would be related either to the field of lexicography or to natural language processing, artificial intelligence etc.
Obligations: Providing feedback on lexicographic practices, workflows and tools

- Observers are expected to participate in ELEXIS project activities dedicated to collecting information about lexicographic practices, workflows and tools.
- These activities include filling in survey questionnaires, or providing other kind of feedback about lexicographic work.
A Survey of Lexicographers' Needs for Institutions

• The aim of the survey is to get an overview of **lexicographic practices** across Europe and worldwide both for born-digital and retrodigitised resources, and to make an inventory of the needs of lexicographers.

• **One survey** needs to be completed **per institution**. (The institution should determine the representative who should complete the survey.)

• He/she might be a **senior lexicographer** or a **computational lexicographer/linguist**. (The expertise of a computational linguist or IT specialist will most likely be required to answer some of the questions.)

• Filling out the survey will take approximately **45 minutes to one hour**.
ELEXIS 2019 event

- Location: Vienna, Austria
- Time: end of February 2019
- Organised by: Austrian Academy of Sciences (as ELEXIS partner)
- Dedicated specifically to ELEXIS observers
  - All observers will be invited
  - The event will be publicized and open to other interested parties
  - Observers will have the opportunity to shape the program (through internal communication channels)
  - It will be an opportunity to discuss ELEXIS post-project sustainability
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.
International Alliance for Lexicography

• 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 organisation dedicated to the advancement of lexicography in the digital age, with the ELEXIS International Advisory Board serving as the **initial governing body**.

• 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.
International Advisory Board

• “ELEXIS consortium partners will be advised by an International Advisory Board.”

• “Members of International Advisory Board will 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.”

• Experts (9) were included in the proposal and Grant Agreement.

• Representatives of continental societies and associations (5) were appointed through Globalex.
Globalex (globalex.link)

• Globalex is a constellation of people who are involved, professionally or academically, with the making of dictionaries and other language resources.
  • The initiative was launched at the eLex 2015 conference in England and includes representatives from all the existing continental associations of lexicography in Africa (Afrilex), Asia (Asialex), Australasia (Australex), Europe (Euralex) and North America (DSNA).
  • Globalex aims to facilitate knowledge sharing and cooperation among its members and with others concerned with language and language technology; promote the creation, research, exchange, dissemination, integration and usage of lexicographical resources and solutions; and enhance interoperability between academia and industry worldwide.
Knowledge sharing (er.elex.is)

• From the Globalex board minutes, 11 January 2018:
  • “Simon is hoping to set up a repository of publications by May. A powerful data mining engine would be used, available on the Globalex website, financed by ELEXIS.”
ELEXIS at Euralex

• Opportunity to discuss your involvement in ELEXIS at Euralex congress:

• ELEXIS booth in the Lobby
• ELEXIS poster session (Grand Union Hall) – Thursday, 16:00-17:00

• info@elex.is
• simon.krek@ijs.si