STSM Scientific report

COST STSM Reference Number:	COST-STSM-IS1305-35156
Period:	2017-01-08 to 2017-01-14 (5 working days)
COST Action:	IS1305
STSM type:	Regular (from Slovenia to Netherlands)
STSM Applicant:	Dr Simon Krek, Jožef Stefan Institute Ljubljana (SI)
STSM Topic:	Matrix dictionary in ELEXIS
Host:	Carole Tiberius, Dutch Language Institute Leiden (NL)

1. Purpose of the STSM

The main goal of the proposed STSM was to elaborate the idea of a matrix dictionary as envisaged in the ELEXIS (European Lexicographic Infrastructure) proposal. ELEXIS proposal aims at linking, integrating and enriching non-integrated modern and historical lexicographic resources available as isolated incompatible data. The goal of ELEXIS is to create a universal (integrated and enriched) registry/network of semantic relations used as a semantic intermediary language for global knowledge exchange, focused on difficult polysemous general vocabulary (single-word and multi-word), modern and historical. Ultimately, this should result in the the realisation of a universal lexicographic metastructure or a »matrix dictionary« spanning across languages and time.

2. Description of the work carried out during the STSM

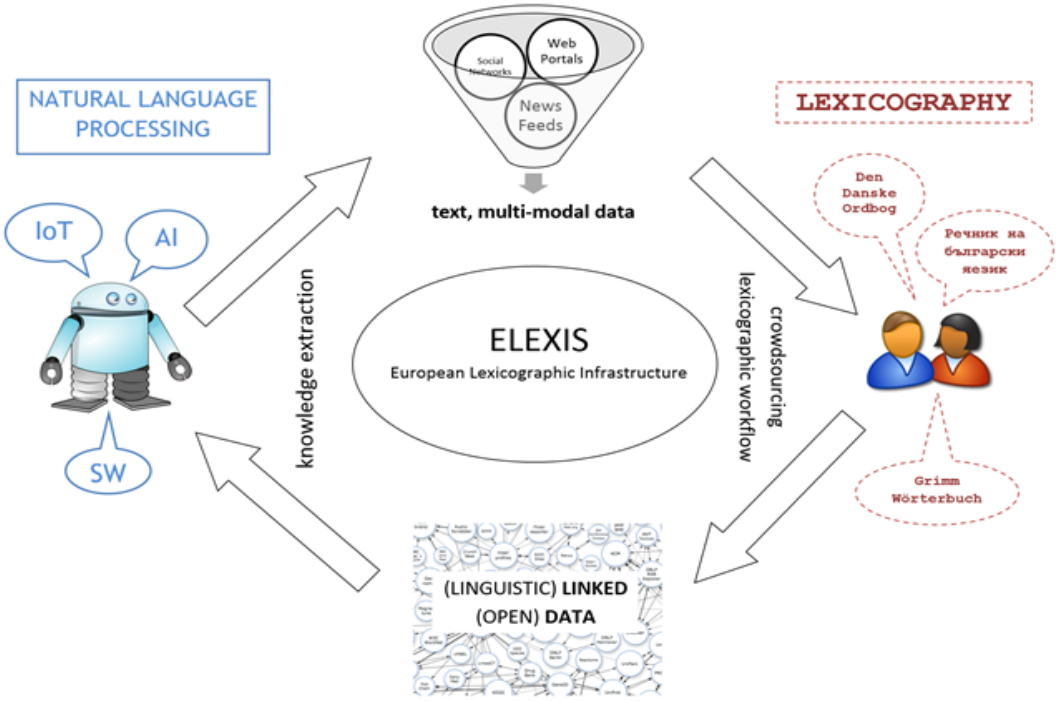
During the STSM the basic concept of the matrix dictionary and the ELEXIS proposal were discussed in detail. To kick-start discussions with INT staff, a presentation was scheduled on Tuesday morning outlining the ELEXIS proposal and the virtuous cycle of elexicography.

ELEXIS will bring 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 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 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 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, 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 rarely 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 proposed infrastructure.



By building a common platform, the so-called virtuous cycle of elexicography can be realised. This means that lexicographic resources will harmonised and integrated into Linked Open Data, making them available to the Natural Language Processing Community. The availability of high-quality semantic information will help NLP applications to achieve higher precision and computational efficiency, which again will support novel lexicography by providing lexicographers with better tools to create new resources.

As a result, stand-alone modern and historical lexicographic resources available as isolated incompatible data will be linked, integrated and enriched on different levels. A scalable, multilingual and multifunctional, language resource will be created by:

- **linking resources**: this means providing links between different elements of dictionary entries (lemmas/headwords, senses, definitions, multi-word expressions, etymologies, etc.) enabling any dictionary (element) to be linked with all other dictionaries (or dictionary elements).
  Result: a growing network of existing dictionaries linked across common concepts via a huge (multilingual) index.

- **integrating resources**: this means taking information from individual resources and putting them together in a new resource / aligning them to create a combined resource.
  Result: any combination of existing (linked) resources resulting in a new resource available for immediate use or as a starting point for creating a novel individual lexicographic resource.

- **enriching resources** with multimodal data (image, sound, video), and unstructured text (corpora, news feeds, social media etc.)
  Result: a portal with cross-lingual, cross-media information on word usage.

Ultimate goal is the creation of a universal (integrated and enriched) 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 realisation of a universal lexicographic metastructure; a matrix dictionary spanning across languages and time.

During the STSM, a lot of time was spent on discussing the concept of linking. The discussion was about the question how to provide links between different elements of dictionary entries (lemmas/headwords, senses, definitions, multi-word expressions, etymologies, etc.) enabling any dictionary (element) to be linked with all other dictionaries (or dictionary elements).
Looking at data from the Slovene Lexical Database and the Dutch ANW (Algemeen Nederlands Woordenboek) and the WNT (Woordenboek der Nederlandse Taal), we concluded that in a first instance, it would be best to focus on the semantic information in the lexicographic resources, rather than to try to convert the whole microstructure of each individual resource into Linked Open Data.
The possibility of linking through a common repository of senses (eg. BabelNet) or directly without an intermediary was also discussed. It also became clear that lexicographic data can play an important role in the further development of common standards (e.g. lemon-ontolex).

The STSM provided also more insight in the requirements for the envisaged ELEXIS infrastructure.

ELEXIS will result in a platform which will consist of several sets of tools and services, as well as new data, in three distinct parts of the platform, or infrastructures.

LEX1: The first set of services and tools will be 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. **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. 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 successful development of this segment of the platform. 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 linking of ELEXIS lexicographic resources will be developed and made available in the **ELEXIS 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, will be 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. Data from this new resource will be available through **ELEXIS matrix dictionary** RESTful Web service as part of the platform.

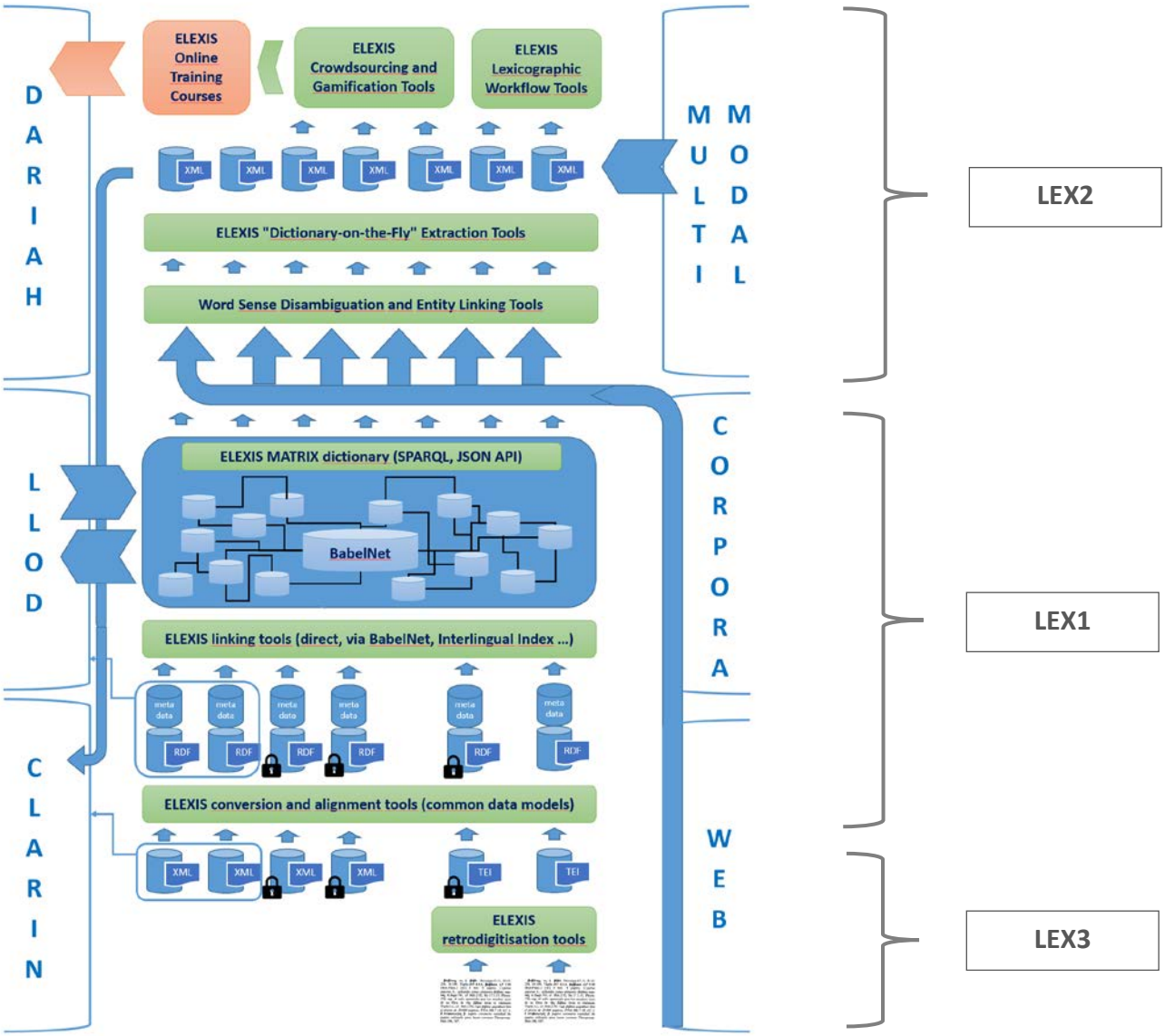
LEX2: Based on the the contribution of lexicographic data in addressing the "semantic bottleneck" a new infastructure will be developed that will include **ELEXIS word sense disambiguation and entity linking tools** dedicated to semantic processing of corpus data, but including tasks such as sense clustering, domain labeling of text, diachronic distribution of senses (ranking senses by frequency of use over time) and similar. 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 **ELEXIS »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.

To enable online lexicographic work on both existing and new (extracted) lexicographic data, two complementary sets of tools will be provided: **ELEXIS lexicographic workflow tools** and **ELEXIS crowdsourcing and gamification tools**. The first will include a user-friendly open

source online dictionary writing system, with the aim to provide the central dictionary writing platform for new lexicography which also includes new possibilities of online collaboration. The other will provide tools for new techniques of dictionary creation, such as explicit or implicit crowdsourcing (gamification). Although similar tools are already in existence (Wiktionary, Urban Dictionary etc.), in ELEXIS the emphasis is on the inclusion of crowdsourcing and gamification techniques in the production of all types of lexicographic resources, also those that are traditionally considered as created exclusively by language description experts.

LEX3: The third set of services is dedicated to retrodigitised dictionaries in **ELEXIS retrodigitisation tools** part of the platform that will include (1) tools for automatic segmentation and structuring of content in retro-digitised dictionaries, and (2) an online generic, modular dictionary publication tool for retrodigitised dictionaries which also offers interfaces for the analysis and profiling of the underlying lexical data. The viewer will include different visualisation, geolocation and profiling tools that make it possible for end users to explore and navigate the dictionary content in novel ways that go beyond the dominant look-up paradigm.

Graphic representation of the ELEXIS platform architecture:



Picture 3 ELEXIS platform

Tools and services: **green,** (lexicographic) data: **blue**, online training and education: **brown,** virtual access infrastructures: **grey**.

3. Conclusion

The STSM was productive. The results of the STSM will feed into the 2nd phase of the ELEXIS grant application and the work resulting from the STSM will be mainly visible in the proposal which is at this point in preparation. Deadline for submission is 29 March 2017.