D.9.2 Report on trans-national access – year 2

Author(s): Sussi Olsen

Bolette S. Pedersen

Date: 31-07-2020

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D. 9.2 Report on trans-national access – year 2

Deliverable Number:	9.2
Dissemination Level:	PU
Delivery Date:	31-07-2020
Version:	V1.0
Author(s):	Sussi Olsen, Bolette S. Pedersen

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Gran Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
V0.1 08-07-2020	First draft	Sussi Olsen, Bolette S. Pedersen
V0.2 16-07-2020	First draft and review	Iztok Kosem
V1.0 31-07-2020	Final version	Sussi Olsen, Bolette S. Pedersen

Report on trans-national access Year 2

Table of Contents

1 Introduction: Trans-national access – Year 2

This deliverable presents the results of year two of the transnational access program of the ELEXIS project. Four visits from the 2nd call were completed in year two. The fifth visit from the 2nd call that had been postponed to spring 2020 has now been postponed again due to the COVID-19 pandemic.

The 3rd call was launched in summer 2019, and the 4th call was launched in winter 2019. Overall, 13 visits have been granted during the second year but due to the COVID-19 crisis only three visits from the 3rd call have actually taken place physically, while one visit was carried out online. The last three visits from the 3rd call and all the visits from the 4th call have been postponed or are awaiting confirmation from the researchers and host institutions, which depends on the development of the COVID-19 situation.

In the following sections, we present the results of the 2nd and 3rd calls for applications, the grant holders origin and projects. Secondly, we examine the profiles of all the grant holders, i.e. gender and experience. Furthermore, we describe the problems for the transnational activities caused by the Covid-19 crisis. Finally, in Annex A, we provide the scientific reports of four of the grant holders from the 2nd call and of the finished visits of the 3rd call, in which the grant holders document their projects and the results of their research visits.

In this report, we do not describe the objectives of the transnational activities of ELEXIS, the format of the visits, the call text and the dissemination, the reviewing process etc. For a thorough description of this, we refer to Deliverable 9.1 from 2019.

2 Calls and grant winners, year 2

During Year 2, two calls for applications were launched, the 3rd call on June 3 2019, and the 4th call on December 11 2020.

The call text was almost the same as in the first two calls , however the emphasis was put on the relevance of contacting the requested hosting institution in advance and providing motivation for the choice of institution in the project description.

The 3rd call received 13 applications, of which the review committee selected seven due to the high quality of the applications.

The seven applicants who were awarded a grant come from Albania, United Kingdom, Spain, Latvia, Israel, and Republic of North Macedonia, and they have visited or will be visiting Denmark, Spain, Austria, Hungary, Serbia, and Slovenia. Table 1 shows the home institutions of the grant winners, the hosting institutions and the project titles.

Home institution	Hosting institution	Project
Faculty of Foreign Languages, University of Tirana	Det Danske Sprog- og Litteraturselskab, University of Copenhagen (DSL/UCPH, Denmark)	A Corpus-based method for Extraction of Polylexical Units (in French and Albanian languages)
Cardiff University	Real Academia Española (RAE, Spain)	Sense Categorization in the Diccionario de la Real Academia Española with Distributional and Lexicographic Supervision
Facultad de Filosofía y Letras (University of Valladolid)	Austrian Academy of Sciences (OEAW, Austria)	Description, creation and exploitation of online lexicographic and terminological resources for the teaching of English Languages for Specific Purposes
Ventspils University of Applied Sciences, Latvia	Austrian Academy of Sciences (OEAW, Austria)	German-Latvian LSP Glossary of Kawall’s “Dieva radījumi pasaulē” and its Original Work
Department of Translation and Interpreting, University of Granada	Hungarian Academy of Sciences (RILMTA, Hungary)	Enhancing EcoLexicon with a phraseological module using the methodology behind Verb Argument Browser
Haifa University	Belgrade Center for Digital Humanities (BCDH, Serbia)	Exploring Digitization and Encoding Options for Ben Yehuda’s Hebrew
Macedonian Academy of Sciences and Arts	Institut Jozef Stefan (JSI, Slovenia)	Creating and maintaining a monolingual Macedonian corpus

Table 1: Home institutions, hosting institutions and projects of the winners of call 3.

The 4th call received 10 applications, of which six were selected. The six grant holders come from Slovakia, Croatia, Spain, Poland, and The Faroe Islands; two of them are going to visit Slovenia, and the others are going to Austria, The Netherlands, Denmark and Estonia. See Table 2 for an overview of the home institutions, the hosting institutions and the projects of the winners of the 4th call.

Home institution	Hosting institution	Project
Ľudovít Štúr Institute of Linguistics, Slovak Academy of Science	Institut Jozef Stefan (JSI, Slovenia)	Dictionary of the Contemporary Slovak Language (DCSL) converted into XML format according to the TEI standard.
Faculty of Teacher Education, University of Zagreb	Institut Jozef Stefan (JSI, Slovenia)	Adapting dictionary writing systems and other platforms to online dictionaries of idioms
University of Granada	Austrian Academy of Sciences (OEAW, Austria)	eLexicography: enhancing the representation of specialized phraseology
Institute of Polish Language, Polish Academy of Science	Institute for Dutch Language (INT, The Netherlands)	INTEGRATION OF LEXICOGRAPHIC DATA: THE DIACHRONIC PLANE
Grunnurin Føroysk Teldutala ('The Faroese Language Technology Foundation'), Tórshavn	Det Danske Sprog- og Litteraturselskab, University of Copenhagen (DSL/UCPH, Denmark)	Ravnur – the Faroese Speech Recognizer
University of Zagreb, Faculty of Humanities and Social Sciences	Institute for Estonian Language (EKI, Estonia)	Automatic detection of neologisms and predictions of their later acceptance

Table 2: Home institutions, hosting institutions and projects of the winners of call 4.

Higher popularity of certain hosting institutions means that some institutions will be unavailable for further visits due to the fixed budget for travel grants given to each hosting institution. As can be seen from Figure 1, this concerns five of the 11 hosting institutions. They have had three or four visits and have spent their budget.

To overcome the future scarcity of hosting institutions available, the project plans on inviting ELEXIS observing institutions (with lexicographic data) to join in the network of ELEXIS hosting infrastructures. Visits to these infrastructures will be without compensation for work carried out by the institution while the visitors will be compensated in the same manner as when visiting existing ELEXIS infrastructures. We foresee that broadening up the infrastructure to include also the observing institutions could benefit the community aspect of the project by including more competences and by providing more flexibility also in relation to the COVID-19 crisis where some countries are in harder difficulties and have more restrictions than others.

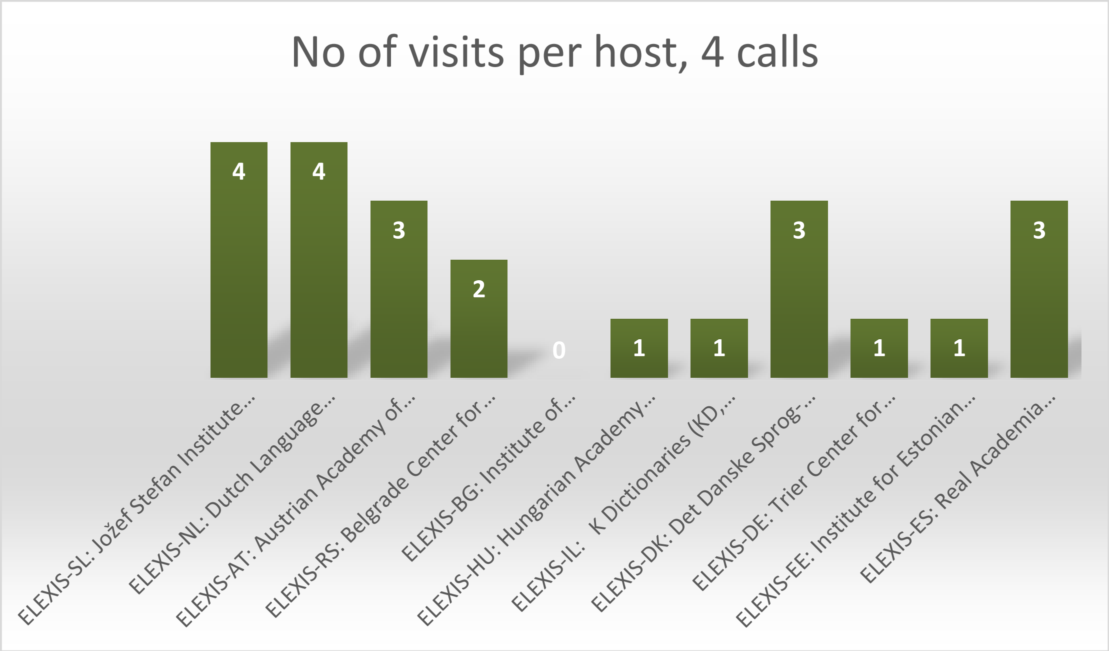


Figure 1: Number of granted visits per host after four calls

3 The COVID-19 situation

The COVID-19 pandemic has been an obstacle to the transnational activities. Several measures have been taken to minimize the impact of the crisis.

As mentioned before, several research visits have been postponed or have not been confirmed by either researcher or the host due to the crisis. One visit was carried out virtually due to time limitations and the grant holder had a fruitful ‘visit’, however, the majority of the grant holders need to meet the hosts physically, e.g. to get access to data not otherwise available or/and to learn from the hosts on the spot. All the hosts as well as grant holders have been very flexible and are awaiting how the situation will evolve.

Apart from postponing the visits, the 5th call that should have been launched in June has been postponed to autumn 2020 in hope that the situation will improve.

The postponement of the research visits will make the schedule tight for some of the hosting institutions, but with the six-month extension that has been approved for the project, it is still possible to accomplish the scheduled amount of research visits if the situation allows.

4 Profiles of grant holders

After four calls, it is now possible to examine more in depth the profile of the applicants and the winners of grants.

We have looked at the gender of the applicants in total as well as of the ones who have been awarded a grant. As can be seen in Figure 2, we received by far more applications from women than from men. Also the succes rate is a little higher for women than for men.

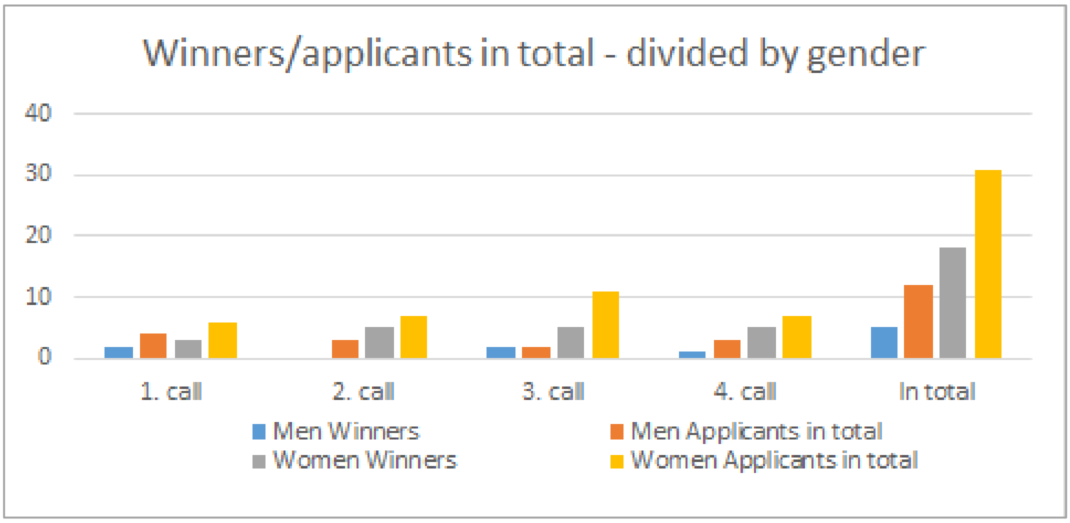


Figure 2: Applicants and winners – gender distribution

Another interesting issue concerns the experience of the grant holders. One might expect that the grant holders would primarily be young researchers or lexicographers applying for a research visit to support their new career. However, as can be seen from Figure 3, this is not the case.

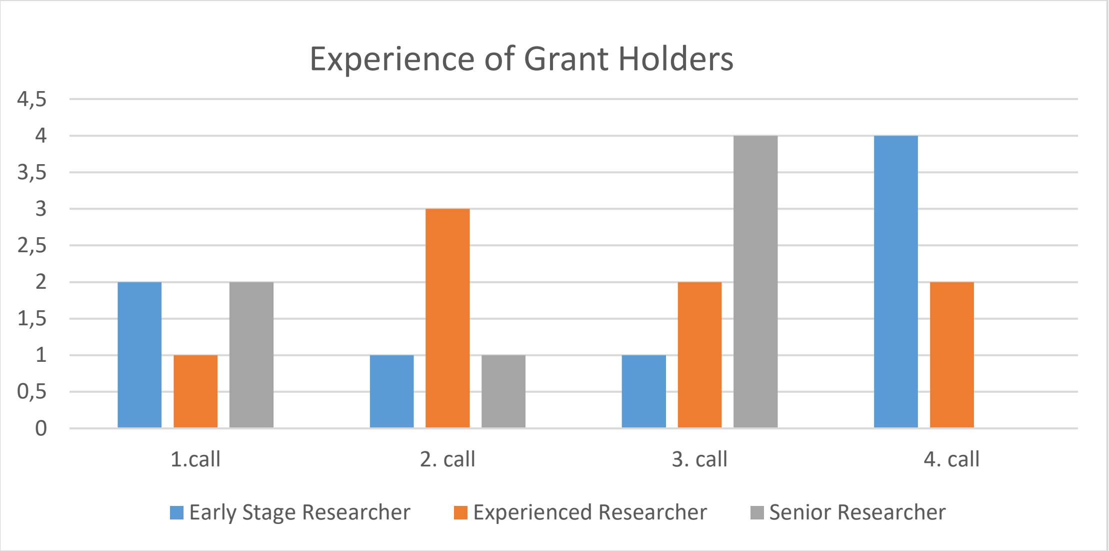


Figure 3: Experience of the grant holders

We divided the grant holders into early stage researchers, experienced researchers, and senior researchers, and the grant holders represent a good mixture of the different levels of experience, a fact that correlates well with the overall mixed goals of the transnational activities. Not surprisingly, most of the experienced researchers apply for projects regarding an upgrade in lexicographical, terminological or technical areas, which are not previously within their experience or which needs upgrade or specialisation. In fact, most grant holders prove to have limited previous experience in the fields that they work with during their research visits.

5 Scientific reports from the second and third call

The reports from the 2nd call visits and from the 3rd call visits that have taken place so far are attached in Appendix A. The grant holders (and hosts) overall reported that the performed visits had been very fruitful in the sense that they had been introduced to relevant resources, tools, projects and working methods of the visited infrastructures. They all reported that their projects have benefited greatly from the visits.

The project topics span from digitization of historic dictionaries or various ways to handle terminology in dictionaries to creating new dictionaries, learning concept representation with neural networks or studying how to present sensitive content in a contemporary dictionary.

6 Concluding remarks on the outcome of year 2

Overall, the second year has been successful as far as the execution of the second and third round of research visits. All grant holders who finished their visits report on high scientific and technical value of these visits.

The COVID-19 pandemic has been an obstacle to the transnational activities in spring 2020. However, alternative solutions such as virtual visits, postponement of visits and postponement of the next call, together with the application for an extension of the project are the measures taken to minimize the impact of the crisis.

As evidenced by the reports, the host institutions have provided the necessary guidance and support during the visits, both scientifically and technically, as to help the grant holders move forward in their projects. Also evidenced by the reports of the grant holders is the fact that the travel grants serve several purposes; first and foremost, the holders gain new knowledge by physically visiting lexicographical milieus with specific expertise in certain for them relevant topics and technologies. Moreover, network building and knowledge exchange are important. And for the individual visiting researchers, the visits serve as a career boost either by helping the early stage researchers establish themselves in the field, or by leading the more experienced researchers towards new fields. The fact that many experienced researchers apply for a research visit to strengthen their knowledge and to gain new expertise indicates that the travel grants meet an existing need not covered by other initiatives.

Most presumably, the grant applications of the upcoming calls will follow the same line as the preceding ones. However, we expect to see an increased interest in the projects of the applicants in the integrative use of the lexicographical tools, methodologies and resources that are currently being or have been developed and made available through ELEXIS, e.g. ELEXIFIER, LEXONOMY, NAISC, Elexifinder, etc.

The project will be presenting the midterm results of the transnational activities at the CLARIN Annual Conference 2020 in October. We are looking forward to this occasion to disseminate the activities to a broader audience and to discuss the value and benefits of the research visits also in relation to the CLARIN initiative.

Appendix A

Scientific reports from the 2nd and 3rdcall are attached below.

Daria Lazić, PhD Student

Institute of Croatian Language and Linguistics

Zagreb, Croatia

Report on Elexis Transnational Research Visit Grant at Det Danske Sprog- og Litteraturselskab and University of Copenhagen

(Copenhagen, Denmark, June 16 – July 5, 2019)

Project title:

Nordic E-dictionaries in Comparison to the *Croatian Web Dictionary – Mrežnik*

Introduction

I applied for a visit to Det Danske Sprog- og Litteraturselskab and the University of Copenhagen with two main goals in mind. Firstly, since I am working on a project within which a new Croatian web dictionary is being compiled, I wanted to visit an institution where a similar project is being conducted in order to exchange ideas and discuss possible common issues. In that respect, Det Danske Sprog- og Litteraturselskab (DSL) was interesting because of *Den Danske Ordbog*, a comprehensive corpus-based dictionary of modern Danish. Secondly, I sought to conduct an initial research for my PhD thesis, in which I will be comparing the work on the Croatian web dictionary with similar foreign dictionaries, in particular dictionaries of Nordic languages. Among other things, I have been interested in the way sensitive content, such as gender, religion, nationality and other, is represented in corpora and dealt with in dictionaries.

The visit has been useful and informative in all the aspects mentioned above. I was introduced to DSL's staff and the projects they are working on and later I had meetings with several editors of *Den Danske Ordbog*. Besides that, everyone was very helpful and eager to answer my questions. I was provided with research material on the topics I am interested in and had an opportunity to explore DSL's tools and resources on my own.

Furthermore, I visited the Centre for Language Technology at the University of Copenhagen, met its researchers and was introduced to their projects, tools and resources. At the end of the visit, I presented the Croatian dictionary project I am working on to the editors of *Den Danske Ordbog*.

In addition, it proved to be a good time to visit due to several events relevant for my work that took place at DSL during my stay there, such as a visit of the lexicographers from the Centre for Digital Lexicography of the German Language and a guest lecture on ethics in lexicography.

Below I will discuss some of the activities mentioned above in more detail.

Scheduled meetings with editors of *Den Danske Ordbog*

During my visit, I met with several of the editors of *Den Danske Ordbog* and discussed diverse topics with them. Some of these are the following:

- introduction to DSL and an overview of their **projects and resources**, in particular *Den Danske Ordbog* (*DDO*; a dictionary of modern Danish, an ongoing project nowadays published online at ordnet.dk/ddo) and *Den Danske Begrebsordbog* (a Danish thesaurus)
- introduction to the **tools** used by the lexicographers at DSL:
    - corpus query system CoREST
    - Word2Dict tool for lemma selection
    - dictionary writing system iLex

    → I was provided with access to the tools as well as the resources which are being edited in iLex, primarily *DDO*, and I had an opportunity to study them on my own. A feature that I found especially useful was the search function in iLex which enables an easy retrieving of information from the dictionary, such as all entries containing a certain usage label, word or semantic field.
- work on *DDO*:
    - **XML structure of the articles** in iLex and the information they contain

→ an interesting feature is for example that the word senses are equipped with genus proximum and an id-number that they share with other lexical resources developed at DSL

- **lemma selection**: candidates for lemmas that could be included in the dictionary can be found in the CoREST corpus tool:
    - frequency word lists generated from the corpora: it is indicated whether a certain word exists in the dictionary, whether someone is already working on the entry or it is free for a lexicographer to compile it
    - suggestions from users together with their comments
    - Word2Dict – a tool that presents semantically related words and indicates whether each of them exists as a lemma in *DDO*; for the lemmas that have already been included in the dictionary the definitions are shown and in that way the tool assists the lexicographer both in selecting new lemmas and writing consistent definitions of lemmas that are semantically related
- **variants in the dictionary**: the dictionary is corpus-based, which means that orthographical, morphological and other variants sometimes appear in the corpus material; both corpus frequencies and conventions proposed by the Danish Language Council are taken into account and the differences are brought up in the dictionary when relevant; since the dictionary examples are taken from the corpus, variants such as different spelling can appear in them
- **revising the dictionary**: along with expanding the dictionary with new lemmas, the existing ones are revised; some of the elements that are revised are spelling (for example in loanwords), definitions and examples (often pointed out by users as problematic for some reason), and lemma inventory (for example, some words that denote phenomena in society or technology can be outdated); currently, an effort is being made to revise controversial words and expressions such as derogatory words and words related to certain social groups or sensitive topics (nationality, religion, gender, age, disabilities, physical features, etc.)

→ practical solutions in *DDO*: an article can be marked for revision at a later point (for example words in the field of technology); genus proximum allows selection of a certain semantic group for revision; example hierarchy – when a new example is added, the older ones can be downgraded or removed

- **stereotypes and potentially offensive content** in the corpus and in the dictionary: problems regarding the presentation of sensitive content (which is commonly related to minority groups in the society) can appear when selecting lemmas for the dictionary, defining their meanings and usage and exemplifying them; part of the problem is that the corpora dictionaries are based on are not always free from stereotypes and offensive language use nor they include texts that specifically regard minority groups; offensive words and expressions are especially problematic because their inclusion in the dictionary is often perceived by the public as an approval of their use

  → I was introduced to the latest discussions about the representation of minority groups in *Den Danske Ordbog*, to reactions from dictionary users and changes they have resulted in

  - among other things, the use of usage labels and explanations boxes has been discussed

  → I was provided with research articles on the topic as well as internal lists of problematic words compiled at DSL, which will be useful for me when studying the way such content is presented in Nordic and Croatian dictionaries and dealing with it in my own work
- **reactions from the dictionary users**: users can both leave comments under a certain article or contribute by suggesting words and expressions that should be added to the dictionary (*Spordhund* function); a list of recent words suggested by users that have been added to the dictionary is published and in that way users are encouraged to participate
- **other resources** developed at DSL:
  - *Den Danske Begrebsordbog*: I have been introduced to the Danish thesaurus, its structure and content, as well as the work on its extension and the logic behind its integration into *Den Danske Ordbog* in the form of

    related words, *ord i nærheden*; I also had an opportunity to browse through the XML document and the printed version of the dictionary
    - *Svensk-Dansk Ordbog* (*Swedish-English Dictionary*) – printed and online version

Visit to the Centre for Language Technology (University of Copenhagen)

During my stay in Copenhagen, I also visited the Centre for Language Technology (Center for Sprogteknologi) at the University of Copenhagen, where I met with the researches who introduced me to their main projects, among which are the following:

- *DanNet* – the Danish WordNet that has been compiled based on the senses in *Den Danske Ordbog* in collaboration with DSL
    - linking to other resources: the researchers explained and showed me the process of linking *DanNet* to *Princeton WordNet*, a project they are currently working on, as well as WordTies, a multilingual WordNet browser for Nordic and Baltic languages
- *The Danish FrameNet* – based on the Berkeley FrameNet model

Furthermore, the researchers at the Centre for Language Technology offered to train two of their tools – lemmatizer and POS-tagger – for Croatian, and I plan on comparing them with the tools already available for Croatian shortly.

Other activities

Among other activities that took place during my visit, it was very interesting to take part in the visit of lexicographers from the Centre for Digital Lexicography of the German Language (https://www.zentrum-lexikographie.de) and follow a seminar where revising and updating the dictionary and communication with the general public were the two main discussion topics. It was a unique opportunity to get an insight into another lexicographic project and follow a discussion on lexicographic problems and solutions.

Another lucky circumstance has been that I could attend a visiting lecture by Dr. René Rosfort from the University of Copenhagen on ethics in lexicographic work, a topic that is both one of my research interests and has proven to be actively discussed among the editors of the Danish dictionary, often encouraged by its users. The lecture sought to explain the problems behind the description of sensitive words and many interesting examples of such content were mentioned.

Finally, at the end of my visit I presented the *Croatian Web Dictionary – Mrežnik* project I am working on to the editors of *Den Danske Ordbog* and brought up a couple of problems that we are facing in our work. The presentation was followed by an interesting discussion which provided useful feedback for my future work.

Conclusion

Looking back on my research visit, I can say that it succeeded beyond my expectations and I believe that the experience and contacts I have gained from it will be extremely valuable for my future work. Over the course of three weeks spent in Denmark, I met researchers from both Det Danske Sprog- og Litteraturselskab and the Centre for Language Technology at the University of Copenhagen and I was introduced to their work. I gained an overview over Danish lexical resources and an in-depth insight into *Den Danske Ordbog*, a resource most relevant for me. Through informative and inspirational conversations with the editors of the Danish dictionary, I became acquainted with the current discussions and trends in Danish and Nordic lexicography, and I was provided with ideas for my own work.

I would like to use this occasion to thank my hosts – all the DSL staff – for their hospitality and for making me feel at home during my stay in Denmark. Furthermore, I would like to thank all the editors and researchers at both Det Danske Sprog- og Litteraturselskab and the University of Copenhagen for eagerly answering all of my questions and providing me with additional material on the topics I found interesting. I would, in particular, like to express my thanks to Dr. Sanni Nimb, Senior Editor at DSL, for assisting me prior to and during my visit as well as planning my activities, and to Sussi Olsen from the University of Copenhagen for organizing my visit to the Centre for Language Technology and for administrative assistance. Finally, I am grateful to the Elexis project for making my research visit possible.

ELEXIS travel grant - my visit at the Belgrade Center for Digital Humanities

Elina Boeva

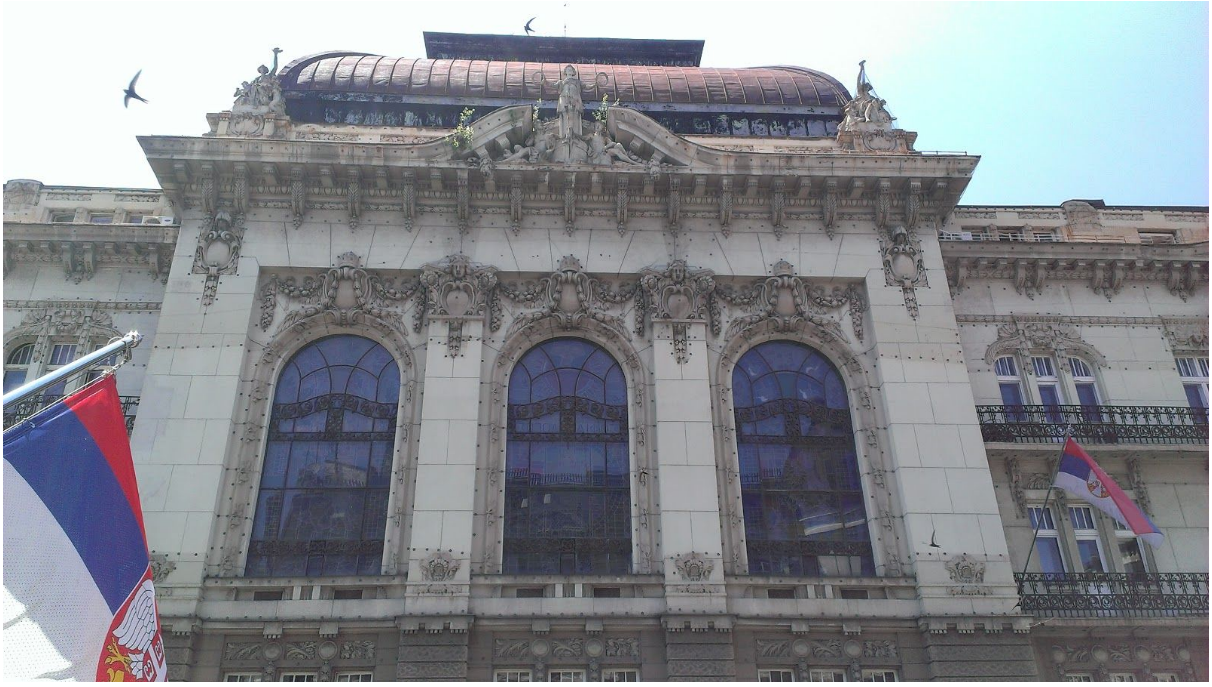
I had the chance to spend the week of June 24-28 in Belgrade exploring the methods for retro-digitizing the *Latin-Bulgarian Dictionary* (ed. M. Voinov, A. Milev) currently in use in academic institutions. This was made possible by ELEXIS - the European lexicographic infrastructure. As an ELEXIS grantholder, I got the opportunity to work with colleagues from the Belgrade Center for Digital Humanities and the Institute for Serbian Language of the Serbian Academy of Arts and Sciences. The visit was supervised by Toma Tasovac, the director of the BCDH, an ELEXIS partner and the leading Serbian institution exploring the use of data modelling, digital editions and standards-compliant lexicographic resources.

1. Travel to Belgrade

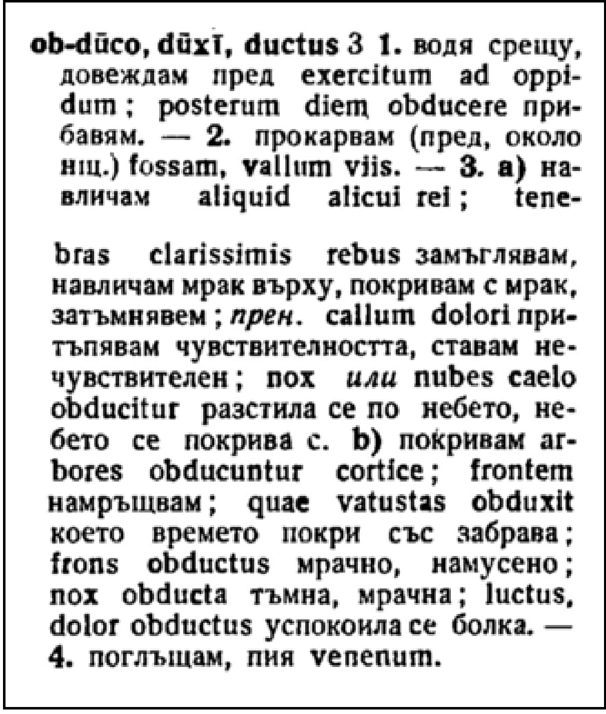
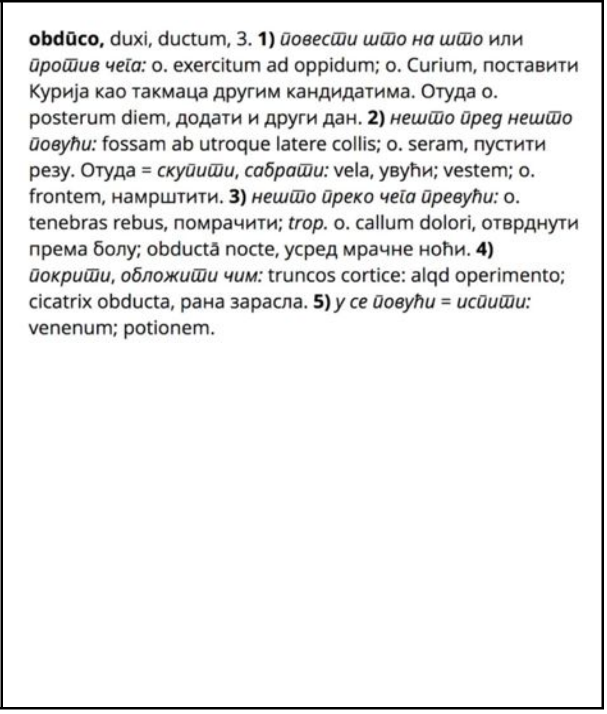
I travelled to Belgrade together with my colleagues Dr Dimitar Iliev (Assistant Professor in Ancient Greek and Latin at Sofia University, National Coordinator for DARIAH-EU in the framework of the CLaDA-BG) and Borislav Petrov, currently a BA student in Classics. The three of us form the core team of a future project for the retro-digitization of our Latin-Bulgarian dictionary, which we hope to be launched within the framework of the CLaDA-BG National Infrastructure in due course.

2. Institute for Serbian language

During the week, we worked in the magnificent building of the BCDH partner Institute for Serbian Language together with our Serbian colleagues and hosts - Toma Tasovac, Marija Gmitrovich, researcher in the Institute and BCDH affiliate, and Justina and Dimitrije, students of Classics at the University of Belgrade. Most of them already have serious experience in digitizing dictionaries. The BCDH and the Institute collaborated on various projects related to the Serbian lexicographic heritage, which lead to the creation of a dictionary platform http://raskovnik.org/ and a platform for the transcription of handwritten heritage http://www.prepis.org/. We got a chance to explore both platforms, learn about the technologies behind them and consider what it would take to adopt similar approaches to creating our own online edition of the Latin-Bulgarian dictionary.



3. The dictionary - structure and inconsistencies.

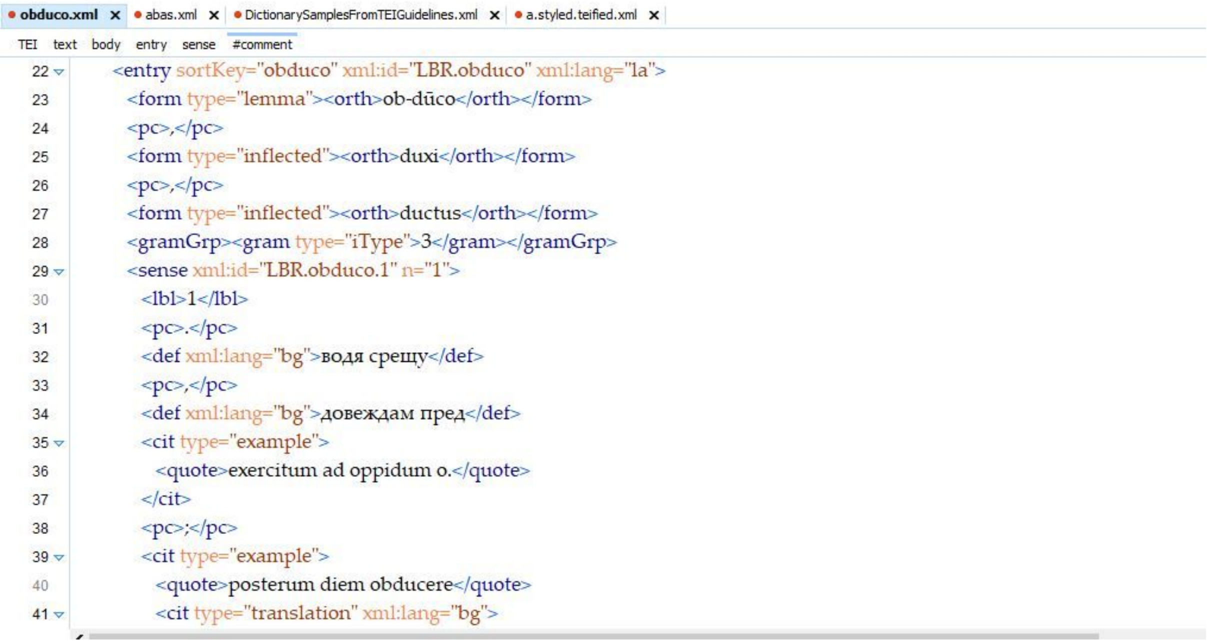
 **ob-dūco, dūxī, ductus** 3 **1.** водя срещу, довеждам пред exercitum ad oppidum; posterum diem obducere прибавям. — **2.** прокарвам (пред, около нщ.) fossam, vallum viis. — **3. a)** навличам aliquid alicui rei; tenebras clarissimis rebus замъглявам, навличам мрак върху, покривам с мрак, затъмнявем; *прен.* callum dolori притъпявам чувствителността, ставам нечувствителен; nox *или* nubes caelo obducitur разстила се по небето, небето се покрива с. **b)** покривам arbores obducuntur cortice; frontem намръщвам; quae vatustas obduxit което времето покри със забрава; frons obductus мрачно, намусено; nox obducta тъмна, мрачна; luctus, dolor obductus успокоила се болка. — **4.** поглъщам, пия venenum.	 **obdūco,** duxi, ductum, 3. **1)** *довести што на што* или *против чега:* o. exercitum ad oppidum; o. Curium, поставити Курија као такмаца другим кандидатима. Отуда o. posterum diem, додати и други дан. **2)** *нешто пред нешто повући:* fossam ab utroque latere collis; o. seram, пустити резу. Отуда = *скупити, сабрати:* vela, увући; vestem; o. frontem, намрштити. **3)** *нешто преко чега превући:* o. tenebras rebus, помрачити; *trop.* o. callum dolori, отврднути према болу; obductā nocte, усред мрачне ноћи. **4)** *покрити, обложити чим:* truncos cortice: alqd operimento; cicatrix obducta, рана зарасла. **5)** *у се повући = испити:* venenum; potionem.

Our initial job was to perform a close analysis of the *Latin-Bulgarian Dictionary*, while our Serbian colleagues worked on a *Latin-Serbian Dictionary* (ed. Jovan Djordjevich). Both dictionaries may have been based on a German source, possibly Wilhelm Freund's *Wörterbuch der Lateinischen Sprache*, which was, after all, also a major source for Lewis and Short as well, after Andrews translated it into English. The structure of the dictionaries is quite similar, but the editors from Bulgaria and from Serbia have made different decisions regarding, for instance, the use examples, abbreviations (e.g. in the Serbian dictionary the abbreviations are in Latin, while in ours they are mostly in Bulgarian) etc.

The first edition of the current Latin-Bulgarian dictionary was published in 1945. It was partially supplemented in the five subsequent re-editions. Although its structure is generally consistent, upon closer look, quite a lot of unexpected “peculiarities” showed up, which made me realise that the digitization is not the only job that is to be done - all the dictionary entries have to be revised and supplemented.

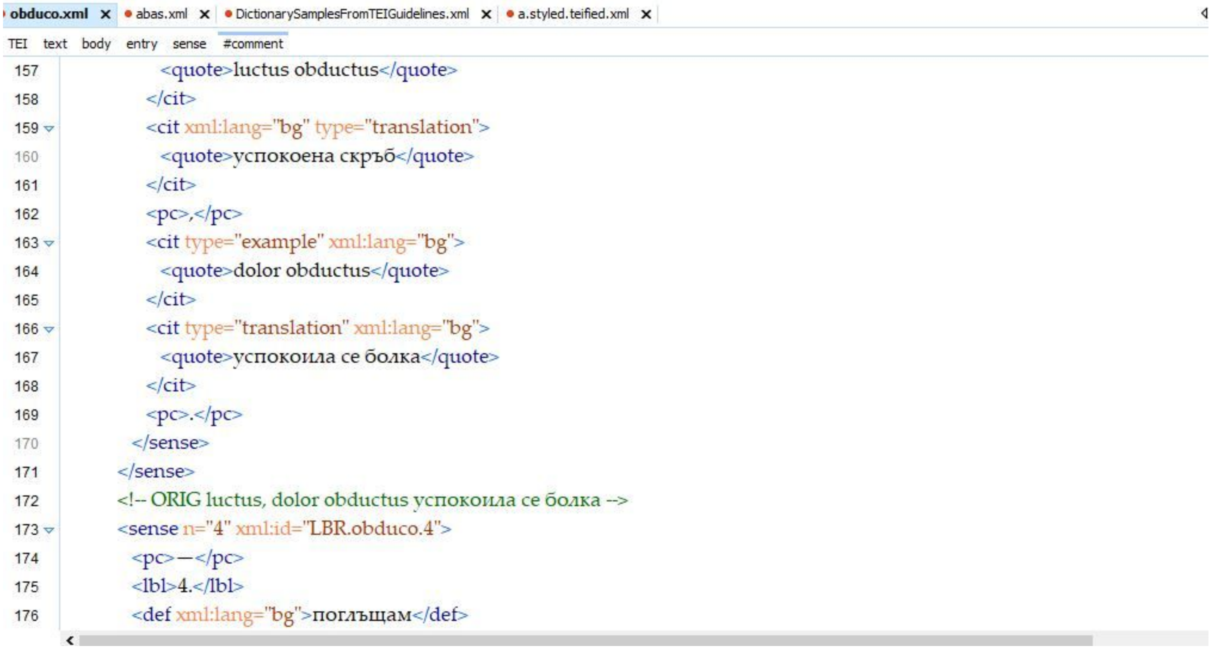
4. Encoding

The next task was to encode our very first lemma under the guidance of Toma Tasovac, who introduced us to the TEI standards and tools, as well as the emerging TEI Lex-0 schema, which is currently under development also in the context of ELEXIS. He “walked” us through the elements, attributes and values used in lexicographic encoding in detail.


obduco.xml
abas.xml
DictionarySamplesFromTEIGuidelines.xml
a.styled.teified.xml
TEI text body entry sense #comment

```
<entry sortKey="obduco" xml:id="LBR.obduco" xml:lang="la">
  <form type="lemma"><orth>ob-dūco</orth></form>
  <pc>,</pc>
  <form type="inflected"><orth>duxi</orth></form>
  <pc>,</pc>
  <form type="inflected"><orth>ductus</orth></form>
  <gramGrp><gram type="iType">3</gram></gramGrp>
  <sense xml:id="LBR.obduco.1" n="1">
    <lbl>1</lbl>
    <pc>.</pc>
    <def xml:lang="bg">водя срещу</def>
    <pc>,</pc>
    <def xml:lang="bg">довеждам пред</def>
    <cit type="example">
      <quote>exercitum ad oppidum o.</quote>
    </cit>
    <pc>;</pc>
    <cit type="example">
      <quote>posterum diem obducere</quote>
      <cit type="translation" xml:lang="bg">
```

During the process of encoding, the need for supplements and changes in the dictionary entries became even more visible, so I decided to encode them all in XML and keep the original form as a comment.


obduco.xml
abas.xml
DictionarySamplesFromTEIGuidelines.xml
a.styled.teified.xml
TEI text body entry sense #comment

```
        <quote>luctus obductus</quote>
      </cit>
      <cit xml:lang="bg" type="translation">
        <quote>успокоена скръб</quote>
      </cit>
      <pc>,</pc>
      <cit type="example" xml:lang="bg">
        <quote>dolor obductus</quote>
      </cit>
      <cit type="translation" xml:lang="bg">
        <quote>успокоила се болка</quote>
      </cit>
      <pc>.</pc>
    </sense>
  </sense>
  <!-- ORIG luctus, dolor obductus успокоила се болка -->
  <sense n="4" xml:id="LBR.obduco.4">
    <pc>—</pc>
    <lbl>4.</lbl>
    <def xml:lang="bg">поглъщам</def>
```

Thanks to Toma Tasovac's detailed introduction, I got acquainted with the TEI guidelines for dictionaries and (as a sort of "homework") was able to encode a few more entries and to explore the differences in encoding a verb, a noun, an adverb and an adjective. The overall approach we chose - with two teams working in parallel on their respective dictionaries while continuously updating each other and comparing each other's work - turned out to be both fun and very productive.

5. Introducing hacks and tricks in Oxygen

One of the most useful things that we learned during our stay in Belgrade were the "hidden possibilities" of Oxygen. I had some experience using Oxygen in encoding epigraphic monuments, but the tips regarding shortcuts, code templates and styling options, for instance, which Toma introduced us to, were completely new to me, and they will greatly optimise not only the future encoding of the *Latin-Bulgarian Dictionary*, but my work in digitization in general.

6. XPath & XSLT

The real challenge was to learn using XPath to navigate through xml-encoded dictionaries. This, thankfully, was made easier by numerous real-life examples we looked at and the exercises that Toma prepared for us, which finally made me feel confident enough to write expressions on my own - and they worked! We were also introduced to the application of some basic XSLT transformation scenarios, that would make our work on the digital Latin-Burlgarian dictionary a lot easier.

7. Cultural program

Though our program was intense and I was completely focused on all the new information, I still managed to appreciate the beauty of Belgrade and visit some of the sights and museums. Belgrade is an unique city and I enjoyed every moment of my stay. We had lots of fun with Toma, Marija, Justina and Nikola (especially laughing at the sound of each-other’s language).

The overall experience with the Belgrade Center for Digital Humanities made me determined to continue and extend my work on the digitization of the Latin-Bulgarian dictionary. I gained an enormous amount of knowledge, that would make the start of the dictionary project much easier. I sincerely hope that we’ll continue working closely with our Serbian colleagues and that their experience will help us in our initial steps towards the edition of the much needed digital Latin-Bulgarian dictionary.

elexis
european lexicographic
infrastructure

CELGA-ILTEC
Centro de Estudos de Linguística Geral
e Aplicada da Universidade de Coimbra

ELEXIS Transnational Research Visit Grant

Final report

Grant holder: Tanara Zingano Kuhn, PhD

Affiliation: Centre for the Study of General and Applied Linguistics at the University of Coimbra, Portugal

Host Institution: Institut Jožef Stefan (JSI, Slovenia)

Host: Iztok Kosem, PhD

Period: 25/6/2019 to 10/7/2019

Project Title: Improving a procedure for automatic extraction of data and import into DWS

1. Purpose of the project

One of the objectives of ELEXIS is to develop tools which can be used by all European institutions working with lexicography in order to promote common standards. One of these tools is the Lexonomy dictionary writing and online dictionary publishing system, which interacts with Sketch Engine for automatic access to corpus data, thus streamlining the process of entry writing and editing.

While in my PhD (Kuhn, 2017) I developed a procedure for automatic extraction of data from a corpus and import into a dictionary writing system (DWS) in order to create a design of a dictionary of Portuguese for university students, my current research focuses on making a prototype for this dictionary, however, this time using Lexonomy. That means that, instead of having to go through a laborious and lengthy two-step process – first extraction of data from the corpus, then import into DWS -, now all this process takes place directly in Lexonomy. This is because, as informed on the ELEXIS website, “Sketch Engine can push lexicographic data into Lexonomy to create automatically-generated dictionary drafts and Lexonomy can pull data from Sketch Engine’s corpora during the entry editing process.”

Thus, it was my purpose during this ELEXIS transnational research visit to contribute to the development of Lexonomy by working on entry modelling. As part of a multilingual team of lexicographers, I was responsible for bringing up some characteristics of the Portuguese language. More specifically, I focused on special lexicographic needs that derive from a perspective of Portuguese as a pluricentric language. My research questions were: how can Portuguese as a pluricentric language be catered for in a model entry? What is necessary to link VOC (a very rich lexical database of Portuguese) with my dictionary prototype on Lexonomy?

One of the goals of this team of lexicographers is to propose a Lexonomy model entry that is as comprehensive as possible so that dictionary makers of any languages can use Lexonomy for their projects. In my research visit, I worked towards helping them achieve that.

2. Description of work carried out during the Research Visit

Firstly, I was introduced to the Lexonomy project in more detail so that I could fully understand the extent to which I could contribute, as well as tune in with their needs. I learned about a very innovative lexicographic project that is currently being developed for Slovene in which Lexonomy is used as DWS. In this project, the functionalities for interaction with Sketch Engine are being successfully adopted, indicating that the tool can be used for the development of other dictionary projects already. In addition to this project, Iztok Kosem gave me an enlightening introduction to a series of other pioneering Slovene lexicography projects that involve, among other state-of-the-art techniques, publication of automatic extracted data in dictionary entries, development of games-with-a-purpose word games and use of crowdsourcing for supporting entry writing. I was thrilled by so many ground-breaking projects and already started conversations with Iztok in order to apply such techniques to the lexicographic projects of Portuguese with which I am involved, such as the development of my dictionary prototype.

One of the paramount conditions for working on entry modelling is to have a common understanding of what each entry element consists and to establish standard terms to refer to them. Thus, I was given access to a shared document regarding literature review on which the above-mentioned group of multilingual lexicographers had been working, which gave me the opportunity to learn from this rich material as well as to contribute to it by doing a series of further specialised readings.

Iztok Kosem and I then had several meetings to discuss what elements should be comprised by the Lexonomy model entry. These were unique moments that gave rise to enlightening theoretical discussions on language, language science and metalexicography. One of our topics of discussion concerned the fact that an entry for a dictionary of Portuguese as a pluricentric language must cater for the fact that there is variation in Portuguese as a result of where it is spoken. I introduced VOC – Vocabulário Ortográfico Comum da Língua Portuguesa (The common spelling dictionary of the Portuguese language) http://voc.cplp.org/ to Iztok Kosem so that he could help me find a solution to link the VOC database to my dictionary prototype in Lexonomy.

3. Description of the main results obtained

At the end of my visit, Iztok and I wrote a model entry in XML that comprised a variety of elements structured in a well-thought-out manner and sent it to the core team of computational linguists working on the development of Lexonomy.

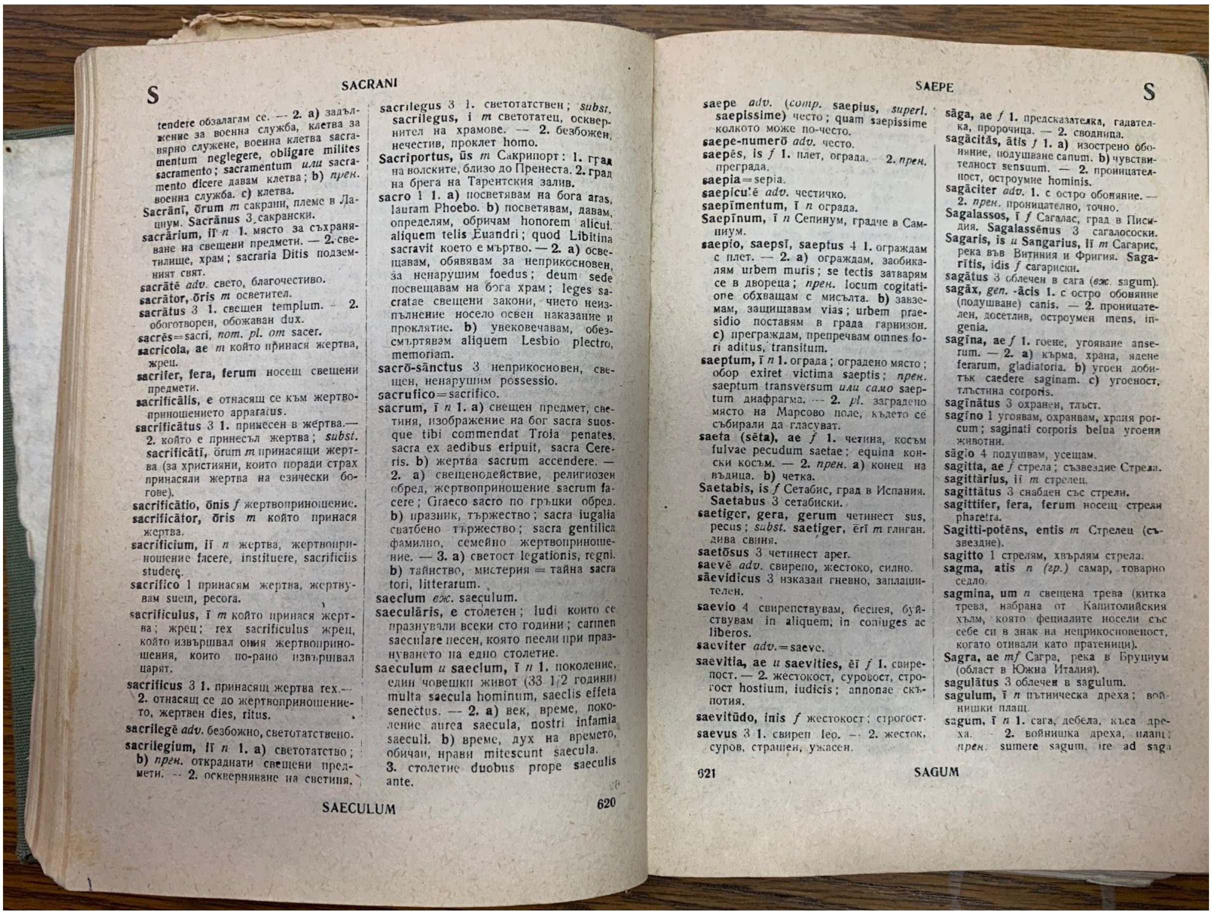
The special case of linking VOC with my dictionary in Lexonomy was carefully debated, however, we could not come up with a final result given lack of time. It is worth mentioning that Iztok showed great interest in this topic, so we will certainly continue research on this issue in a future work.

Even though my contribution is only a small part of a much larger project, I am glad to know that the actual needs concerning the Portuguese language will be considered when creating a model entry for Lexonomy. As a result, I will be able to carry out my project of a dictionary of Portuguese for university students using the most advanced techniques for dictionary making. Moreover, any other lexicographer working on dictionaries of Portuguese will benefit from this ground-breaking tool.

5. Concluding remarks

This Transnational Research Visit Grant has been highly instructive and motivational. I have had the opportunity to learn about pioneering lexicographic projects for Slovene, which not only contributed to knowledge growth, but also served as an incentive for me to share this information with colleagues at my host institution. I am highly motivated to further learn about those techniques in order to work on adjustments for implementation of projects related to Portuguese.

It should be highlighted that those meetings with Iztok Kosem took place in different locations, namely, Institut Jožef Stefan, Trojína Research Centre and University of Ljubljana. Such variety allowed me to visit different laboratories and research institutes and get to know researchers working on a number of lexicographic projects. There is no doubt that this was a great opportunity for expanding my academic network.



S SACRANI

tendere обзалагам се. — **2. a)** задължение за военна служба, клетва за вярно служене, военна клетва sacramentum neglegere, obligare milites sacramento; sacramentum *или* sacramento dicere давам клетва; **b)** *прен.* военна служба. **c)** клетва.

Sacrānī, ōrum *m* сакрани, племе в Лациум. **Sacrānus** 3 сакрански.

sacrārium, iī *n* **1.** място за съхраняване на свещени предмети. — **2.** светилище, храм; sacraria Ditis подземният свят.

sacrātē *adv.* свето, благочестиво.

sacrātor, ōris *m* осветител.

sacrātus 3 **1.** свещен templum. – **2.** обоготворен, обожаван dux.

sacrēs=sacri, *nom. pl. от* sacer.

sacricola, ae *m* който принася жертва, жрец.

sacrifer, fera, ferum носещ свещени предмети.

sacrificālis, e отнасящ се към жертвоприношението apparatus.

sacrificātus 3 **1.** принесен в жертва.— **2.** който е принесъл жертва; *subst.* **sacrificātī, ōrum** *m* принасящи жертва (за християни, които поради страх принасяли жертва на езически богове).

sacrificātio, ōnis *f* жертвоприношение.

sacrificātor, ōris *m* който принася жертва.

sacrificium, iī *n* жертва, жертвоприношение facere, instituere, sacrificiis studere.

sacrifico 1 принасям жертва, жертвувам suem, pecora.

sacrificulus, ī *m* който принася жертва; жрец; rex sacrificulus жрец, който извършвал ония жертвоприношения, които по-рано извършвал царят.

sacrificus 3 **1.** принасящ жертва rex.— **2.** отнасящ се до жертвоприношението, жертвен dies, ritus.

sacrilegē *adv.* безбожно, светотатствено.

sacrilegium, iī *n* **1. a)** светотатство; **b)** *прен.* откраднати свещени предмети. – **2.** оскверняване на светиня.

sacrilegus 3 **1.** светотатствен; *subst.* **sacrilegus, i** *m* светотатец, осквернител на храмове. — **2.** безбожен, нечестив, проклет homo.

Sacriportus, ūs *m* Сакрипорт: **1.** град на волските, близо до Пренеста. **2.** град на брега на Тарентския залив.

sacro 1 **1. a)** посветявам на бога aras, lauram Phoebo. **b)** посветявам, давам, определям, обричам honorem alicui, aliquem telis Euandri; quod Libitina sacravit което е мъртво. — **2. a)** освещавам, обявявам за неприкосновен, за ненарушим foedus; deum sede посвещавам на бога храм; leges sacratae свещени закони, чието неизпълнение носело освен наказание и проклятие. **b)** увековечавам, обезсмъртявам aliquem Lesbio plectro, memoriam.

sacrō-sānctus 3 неприкосновен, свещен, ненарушим possessio.

sacrufico=sacrifico.

sacrum, ī *n* **1. a)** свещен предмет, светиня, изображение на бог sacra suosque tibi commendat Troia penates, sacra ex aedibus eripuit, sacra Cereris. **b)** жертва sacrum accendere. — **2. a)** свещенодействие, религиозен обред, жертвоприношение sacrum facere; Graeco sacro по гръцки обред. **b)** празник, тържество; sacra iugalia сватбено тържество; sacra gentilica фамилно, семейно жертвоприношение. — **3. a)** светост legationis, regni. **b)** тайнство, мистерия = тайна sacra tori, litterarum.

saeclum *вж.* saeculum.

saeculāris, e столетен; ludi които се празнували всеки сто години; carmen saeculare песен, която пеели при празнуването на едно столетие.

saeculum *и* **saeclum, ī** *n* **1.** поколение, един човешки живот (33 1/2 години) multa saecula hominum, saeclis effeta senectus. — **2. a)** век, време, поколение aurea saecula, nostri infamia saeculi. **b)** време, дух на времето, обичаи, нрави mitescunt saecula. **3.** столетие duobus prope saeculis ante.

SAECULUM 620

SAEPE S

saepe *adv.* (*comp.* **saepius**, *superl.* **saepissime**) често; quam saepissime колкото може по-често.

saepe-numerō *adv.* често.

saepēs, is *f* **1.** плет, ограда. – **2.** *прен.* преграда.

saepia=sepia.

saepicu!ē *adv.* честичко.

saepīmentum, ī *n* ограда.

Saepīnum, ī *n* Сепинум, градче в Самниум.

saepio, saepsī, saeptus 4 **1.** ограждам с плет. — **2. a)** ограждам, заобикалям urbem muris; se tectis затварям се в двореца; *прен.* locum cogitatione обхващам с мисълта. **b)** завземам, защищавам vias; urbem praesidio поставям в града гарнизон. **c)** преграждам, препречвам omnes fori aditus, transitum.

saeptum, ī *n* **1.** ограда; оградено място; обор exiret victima saeptis; *прен.* saeptum transversum *или само* saeptum диафрагма. — **2.** *pl.* заградено място на Марсово поле, където се събирали да гласуват.

saeta (sēta), ae *f* **1.** четина, косъм fulvae pecudum saetae; equina конски косъм. — **2.** *прен.* **a)** конец на въдица. **b)** четка.

Saetabis, is *f* Сетабис, град в Испания. **Saetabus** 3 сетабиски.

saetiger, gera, gerum четинест sus, pecus; *subst.* **saetiger, ērī** *m* глиган. дива свиня.

saetōsus 3 четинест aper.

saevē *adv.* свирепо, жестоко, силно.

sāevidicus 3 изказан гневно, заплашителен.

saevio 4 свирепствувам, беснея, буйствувам in aliquem; in coniuges ac liberos.

saeviter *adv.*=saeve.

saevitia, ae *u* **saevities, ēī** *f* **1.** свирепост. — **2.** жестокост, суровост, строгост hostium, iudicis; annonae скъпотия.

saevitūdo, inis *f* жестокост; строгост.

saevus 3 **1.** свиреп leo. — **2.** жесток, суров, страшен, ужасен.

sāga, ae *f* **1.** предсказателка, гадателка, пророчица. — **2.** сводница.

sagācitās, ātis *f* **1. a)** изострено обоняние, подушване canum. **b)** чувствителност sensuum. — **2.** проницателност, остроумие hominis.

sagāciter *adv.* **1.** с остро обоняние. — **2.** *прен.* проницателно, точно.

Sagalassos, ī *f* Сагалас, град в Писидия. **Sagalassēnus** 3 сагалососки.

Sagaris, is *u* **Sangarius, iī** *m* Сагарис, река във Витиния и Фригия. **Sagarītis, idis** *f* сагариски.

sagātus 3 облечен в сага (*вж.* sagum).

sagāx, *gen.* **-ācis** **1.** с остро обоняние (подушване) canis. — **2.** проницателен, досетлив, остроумен mens, ingenia.

sagīna, ae *f* **1.** гоене, угояване anserum. — **2. a)** кърма, храна, ядене ferarum, gladiatoria. **b)** угоен добитък caedere saginam. **c)** угоеност, тлъстина corporis.

sagīnātus 3 охранен, тлъст.

sagīno 1 угоявам, охранвам, храня porcum; saginati corporis belua угоени животни.

sāgio 4 подушвам, усещам.

sagitta, ae *f* стрела; съзвездие Стрела.

sagittārius, iī *m* стрелец.

sagittātus 3 снабден със стрели.

sagittifer, fera, ferum носещ стрели pharetra.

Sagitti-potēns, entis *m* Стрелец (съзвездие).

sagitto 1 стрелям, хвърлям стрела.

sagma, atis *n (гр.)* самар, товарно седло.

sagmina, um *n* свещена трева (китка трева, набрана от Капитолийския хълм, която фециалите носели със себе си в знак на неприкосновеност, когато отивали като пратеници).

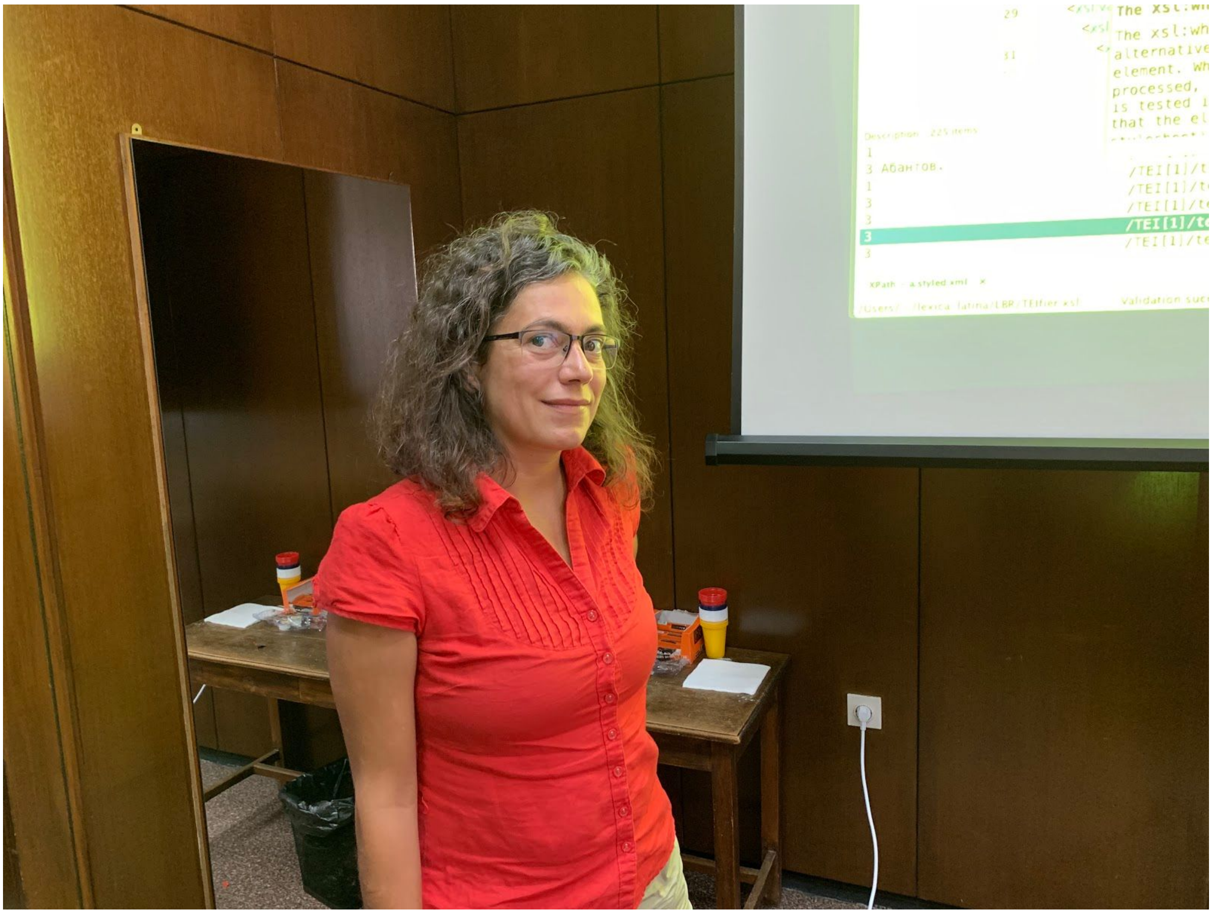
Sagra, ae *mf* Сагра, река в Бруциум (област в Южна Италия).

sagulātus 3 облечен в sagulum.

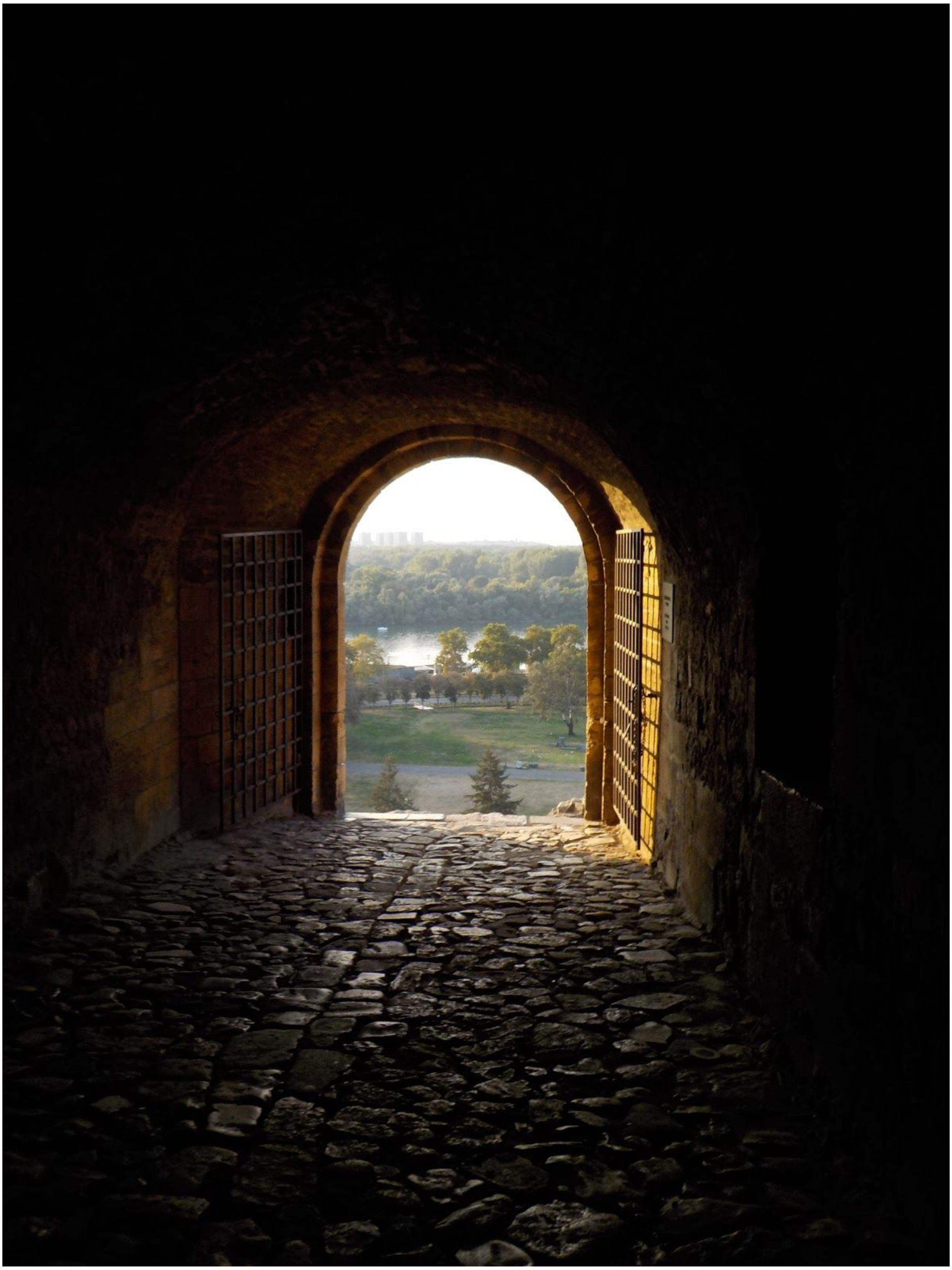
sagulum, ī *n* пътническа дреха; войнишки плащ.

sagum, ī *n* **1.** сага, дебела, къса дреха. – **2.** войнишка дреха, плащ; *прен.* sumere sagum, ire ad saga

621 SAGUM


The xsl:wh
The xsl:wh
alternative
element. Wh
processed,
is tested i
that the el
Description 225 items
3 Абаштов.
/TEI[1]/t
XPath a.styled.xml
Validation suc

acer



ELEXIS REPORT

ELEXIS Transnational Research Visit Grant

Final Report – Anna Tenieshvili,
English-Georgian Maritime Dictionary

Travel Grant: `	***Call 2***
Hoisting Institution:	***Dutch Language Institute***
Period of Stay:	***June 17-28, 2019***
Researcher:	***Anna Tenieshvili***
Affiliation:	***Batumi State Maritime Academy, Batumi, Georgia***
Current Position:	***Associate Professor of Foreign Languages Department at Batumi State Maritime Academy***
Project Title:	***New Modern English-Georgian Maritime Dictionary***

Introduction

I applied to Dutch Language Institute as among all institutes listed on the internet page of ELEXIS Transnational Research Visit Grant Program, this Institute was named as the right institution to instruct me from theoretical, practical and technical points of view for the project of compilation of new English-Georgian Maritime Dictionary I am going to undertake. The essence of my project is to compile new modern English-Georgian Maritime Dictionary comprising translation of maritime terms into Georgian, their definitions, examples from corpus and some additional information. The article titled “Why it is Necessary to Create and Adopt Georgian Maritime Terminology” connected to this project was presented by me at the conference organized by Giorgi Akhvlediani Society for the History of Linguistics “International Terminology: Translation and Standardization” held at Tbilisi Ivane Javakhishvili State University in October of 2018.

Research Goals

In order to compile dictionary that will comprise various information including terms definitions, corpus examples, some encyclopedic information, related words, etc, it was necessary to conduct some research work, to listen to advice of specialists experienced in similar work. As my dictionary is intended to have both terminological and lexicographic features, the opinions of terminologists and lexicographers are important to be taken into consideration when composing entries and compiling the dictionary as a whole. In addition to the dictionary compilation, the purpose of my work is to conduct research in order to fill in the lexical gaps existing in maritime terminology of the Georgian Language. Some useful advice was given to me at Dutch Language Institute that I can share with terminologists in Georgia when work on creation of Georgian Maritime terms is started. The issue of filling in the above-mentioned lexical gaps is very important as full version of English-Georgian Maritime dictionary can be available only after all necessary maritime terms are created in the Georgian language to fill in the lexical gaps existing at present.

Methodological Plan

- to receive information regarding term formation
- to receive information regarding term definitions
- to receive information regarding English/Georgian language corpora and choice of right examples
- to receive information on technical side of the project
- to start compilation of dictionary entries

Research questions

Compliance of term definitions with ISO 704 standard

Usage of Sketch Engine Tools for Corpora Compilation

Usage of QTerm, MultiTems databases to compile terminological databases

Familiarization with dictionary writing systems including LEXONOMY to start dictionary compilation

Receiving instructions regarding term definitions

Receiving instructions regarding entries compilation

Work Process

On the first day of my visit to Dutch Language Institute I gave presentation on my project of compilation of new modern English-Georgian Maritime Dictionary explaining the contemporary situation in the field of Georgian lexicography related to maritime field; the role, Georgia plays on international maritime arena today, substantiating the necessity to compile the new English-Georgian Maritime dictionary. During the first week of my visit to Dutch Language Institute I had very interesting meetings and consultations with the staff of Dutch Language Institute experienced in compilation of monolingual and bilingual dictionaries.

I got introduced to terminological databases, QTerm and MultiTerms, terminologists of Dutch Language Institute explained the principles of terms creation to me, I was explained principle of work with Sketch Engine, principles of lexical work, and difference between terminological and lexicographic approaches of dictionary compilation. I was also introduced to different dictionaries creating databases.

For my project it is important to invest time in compiling a balanced corpus for the specific domain as a lot of documents are not available in digital form or are copyright protected. In addition I am going to contact the library of Maritime Academy I am currently employed for to check availability of relevant materials for my corpus.

During the second week of my visit to Dutch Language Institute, I started compilation of my dictionary, using Internet-based dictionary writing system. For compilation of new modern English-Georgian Maritime Dictionary I am going to use the combined approach that implies processing terminology on basis of lexicographic structures. During this practical work I consulted the compliance of term definitions of my dictionary with ISO standard and corresponding requirements with terminologists at Dutch Language Institute and continued my work on compilation of dictionary entries, having taken their comments and remarks into consideration.

Conclusions

It has been very fruitful to me to have an opportunity to receive advice and consultations of terminologists, lexicographers and other specialists from Dutch Language Institute.

I have received the right direction for my work and kind of guidance that enabled me to start compiling my dictionary already during the second week of my visit to Dutch Language Institute.

The guidance I have received at Dutch Language Institute has simultaneously practical and theoretical character and now when I have started work on compiling new modern English-Georgian Maritime Dictionary I try to follow all instructions from my colleagues at Dutch Language Institute.

I am very grateful to ELEXIS Transnational Research Visit Grant and Dutch Language Institute for having opportunity to visit Dutch Language Institute for two weeks. This visit was very useful to me as it has given me a lot of knowledge of informational, instructional and practical character and what is most important, the visit itself, interaction with specialists of the field, their instructions and guidance inspired me for immediate start of my project.

I have decided to give the dictionary the open access as soon as I have compiled the certain number of entries to facilitate the studies process for the students of maritime field in Georgia and for usage by all interested parties.

As I proceed with my project I am going to present the report on this project in the form of article at EURALEX conference to be held in Greece in 2020.

Appendix

In order to illustrate the entries from new modern English-Georgian Maritime Dictionary, hereby I bring samples of several entries from this dictionary:

anchor n
[ˈæŋ.kər]

ღუზა

appliance now consisting of a heavy iron, composed of a long shank, having a ring at one end to which the cable is fastened, and at the other branching out into two arms or flukes, tending upwards, with barbs or edges on each side; this appliance is intended for holding a ship, etc., fixed in a particular place, by mooring it to the bottom of the sea or river
სპეციალური ფორმის ნაჭედი, სხმული ან შენადუღი კონსტრუქცია, რომლის დანიშნულებაა გემის ან სხვა მცურავი ობიექტის შეკავება ზღვაში გრუნტთან ურთიერთქმედების ხარჯზე

Corpus examples
An undersea crude oil pipeline ruptured on Sunday after being hit by a ship's anchor , spilling over 20,000 gallons of crude oil into the Gulf of Mexico and leaving a half-mile long oil slick on the water.
მარჯვენა მხარე — ოქროს საწმისის გამოსახულებით, გვახსენებს ჩვენ მითს არგონავტებზე, რომლებმაც ჩაუშვეს ღუზა სანაპიროებს, სადაც ეხლა ფოთი მდებარეობს. ქალაქის თანამედროვე გერბის საფუძვლად აღებულია 1858 წლის ქალაქის გერბი.

cargo n
[ˈkɑː.ɡəʊ]

—

ტვირთი

goods carried by a ship, car or an aircraft; in maritime use: freight or lading of a ship, a ship-load
გარკვეული სიმძიმის ბარგი, რომელიც ერთი ადგილიდან მეორეზე გადააქვთ

Corpus examples
Before passengers were permitted to leave, a government official inspected the ship's cargo.
... აკრძალონ გემის შესვლა საქართველოს შიდა საზღვაო წყლებში, ნავსადგურებში და რეიდებზე, თუ გემი ან მისი ტვირთი საფრთხეს უქმნის ადამიანის ჯანმრთელობასა და სიცოცხლეს ან ზღვის ცოცხალ რესურსებს
to cargo
v

keel n
[kiːl]

—

გემის კილი

lowest longitudinal timber of a ship or boat, on which the framework of the whole is built up;
in iron vessels, a combination of iron plates taking the place and serving the purpose of the keel of a wooden vessel
ძირითადი გასწვრივი ფსკერისეული კოჭი გემის დიამეტრულ სიბრტყეში; გემის ფსკერის ძირითადი კოჭი, რომელიც გადებულია მთელ სიგრძეზე (გემის მდგრადობისათვის)

Corpus examples
"Wreck which lay in Dusky Sound was Cook's ship. He made several measurements to prove his point, but as the vessel's keel was 128 feet and her tonnage appeared to be about 800 tons this was unlikely. Moreover he overlooked some important points
წიგნი „დაბადება“ არაფერს ამბობს იმის შესახებ, ჰქონდა თუ არა კიდობანს კილი, ცხვირი, აფრები, ნიჩბები თუ საჭის ფრთები. აღსანიშნავია, რომ კიდობნად ნათარგმნი ებრაული სიტყვა გვხვდება გამოსვლის 2:3, 10-შიც, სადაც ნათქვამია, რომ დედამ პატარა მოსე შეფისულ კალათაში [კიდობანში] ჩააწვინა და მდინარე ნილოსს გაატანა.

Silga, Sviķe, *Dr. philol.*
Ventspils University of Applied Sciences

Report on ELEXIS Transnational Research Visit Grant at the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences (ACDH-OeAW).

(Vienna, Austria March 16-20 2020)

Travel Grant: Call 3

Project title: "German-Latvian LSP Glossary of Kawall's "Dieva radījumi pasaulē" and its Original Work"

Introduction

My visiting grants project proposal is part of a larger project aiming to research Latvian botanical terminology used in H. Kawall's work "God's Creatures in the World" ("Dieva radījumi pasaulē"). This work (a textbook) is one of the first translations from German into Latvian, in which the author mentions Latvian botanical terms for the first time ever, in addition to the terms of zoology and mineralogy, the book has a separate chapter – Plant Kingdom (Augu valsts). A detailed research of botanical lexis requires a digital corpus of language material on which to compare and study special lexis used in the original language and translation. Therefore, the **goal** of the research stay is to create a bilingual digital LSP corpus based on the original book in German and its translation into Latvian, along with the aim to compile a bilingual LSP glossary that includes a collection of special botanical vocabulary used in H. Kawall's translation and original work.

In this report, I will present the workflow of my research visit, i.e. preparatory work, used books, support and main work at the ACDH-OeAW, materials studied and tools used to the achieve the goal – create two corpora and a glossary. I will summarize the conclusions and single out some possible solutions for the further research process.

Workflow and description of steps for performing specific tasks

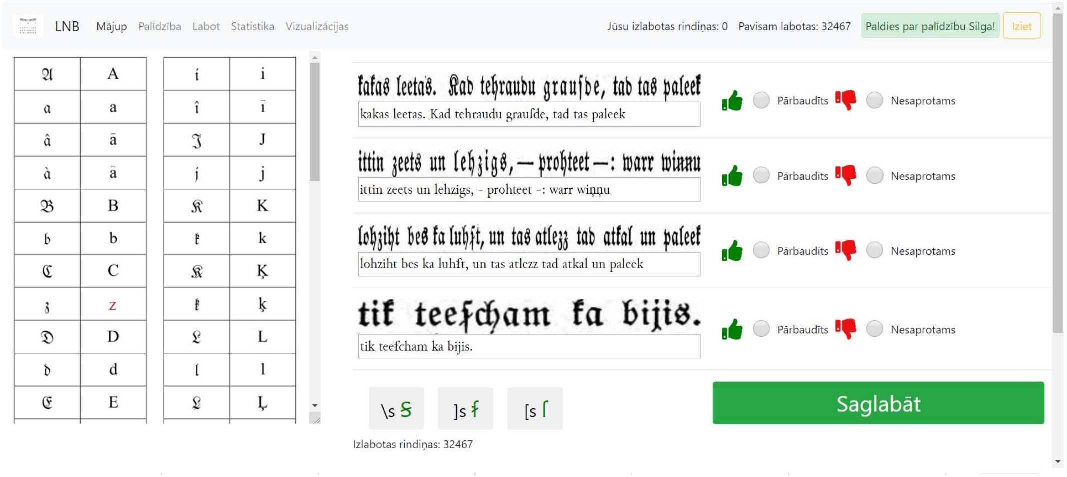
Conducting preparatory work before the research visit.

Preparation of two printed books: H. Kawall "Deewa raddijumi pasaulē" („Dieva radījumi pasaulē" (DRP)) (1860) and "Die Naturgeschichte für Kinder und Elementarschüler, oder erster Unterricht über das Mineralreich, Pflanzenreich und Tierreich, mit über 300 kolorierten Abbildungen" (1855). Scanning and saving both books in PDF format.

Digitising paper books (OCR scans).

For solving theoretical issues, finding tools and methods of digitisation of both scanned books the support of the National Library of Latvia (NLL) was sought. The scanned documents had to be prepared in OCR (optical character recognition) format, i.e. in such a format to make it possible to find, edit and process specific fragments of the text using the search function. The documents could not be pictures, such as jpg format. Doc-Works programme environment was used to process texts by working with the scanned material. The texts of both books are written in the old script, and it was the greatest challenge of the project. The new supervised machine learning approach offered by NLL was used to digitise the texts, and these two books were the first ones to be processed like this. Initially the text was recognised by an untrained algorithm. Next, the text was edited.

For the computer to recognise text accurately, a sample of the correct representation of the text must be made first – at least 10 000 perfect, human-edited lines. Each line must be checked by two writers editing individual lines. You can do it and view the process on the NLL website: https://frakturs.lnb.lv/ (see Fig. 1)



𝔄	A	i	i
𝔞	a	î	ī
â	ā	ℑ	J
à	ā	𝔧	j
𝔅	B	𝔎	K
𝔟	b	𝔨	k
ℭ	C	𝔎	Ķ
𝔷	z	𝔨	ķ
𝔇	D	𝔏	L
𝔡	d	𝔩	l
𝔈	E	𝔏	Ļ

kakas leetas. Kad tehraudu grauſde, tad tas paleek

ittin zeets un lehzigs, - prohteet -: warr wiņņu

lohziht bes ka luhft, un tas atlezz tad atkal un paleek

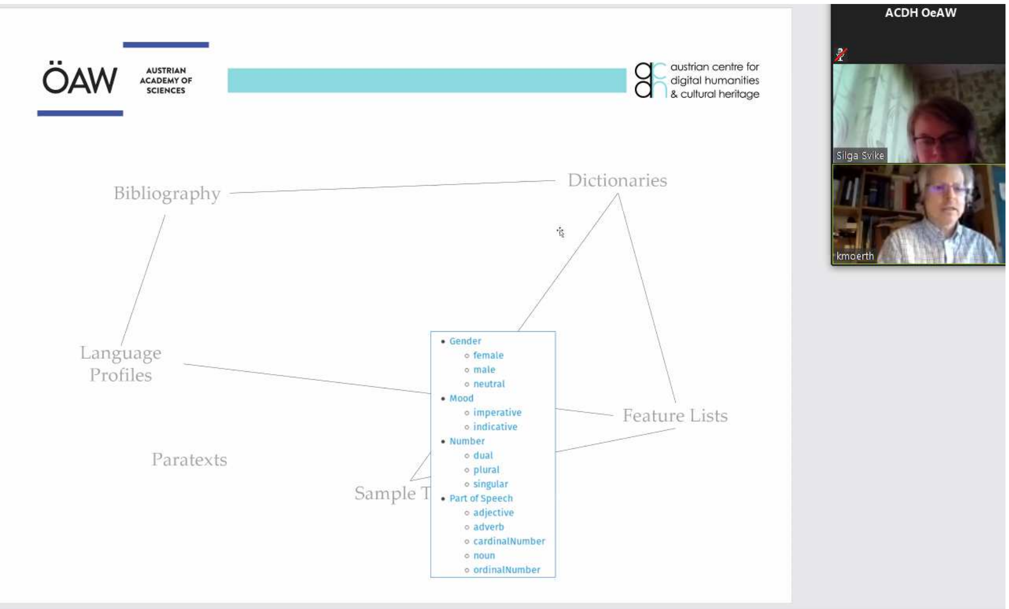
tik teeſcham ka bijis.

Figure 1. Editing Gothic texts manually

To enable automatic recognition, the Tesseract text recognition software was used. It works using the LSTM (long short-term memory) neural network model. LSTM operates more accurately than the early neural network models, and it is well-suited for script and speech recognition. LSTM belongs to the deep learning algorithms. The Tesseract software is a completely new solution of the year 2019, and it was used in this project as an experiment. This program used the German text model as the basic model, where the letters have diacritical marks.

During research visit.

Participation in the virtual meeting with deputy head of the ACDH-OeAW, Dr. Karlheinz Mörth, who shared expertise regarding best practice and standards in lexicography (see Figure 2). During the virtual meeting there was a possibility follow insights into different projects carried out in the ACDH-OeAW e.g. Vienna Corpus of Arabic varieties (VICAV). It was a good opportunity to read the study “Best practices for lexicography – intermediate report” by C. Tiberius, R. Costa, T. Erjavec, S. Krek, J. McCrae, C. Roche, T. Tasovac, 31 January 2020 (available at: http://ejuz.lv/estpractices).


ÖAW
AUSTRIAN ACADEMY OF SCIENCES
austrian centre for digital humanities & cultural heritage
ACDH OeAW
Silga Svike
kmoerth

Bibliography

Dictionaries

Language Profiles

Paratexts

Sample T

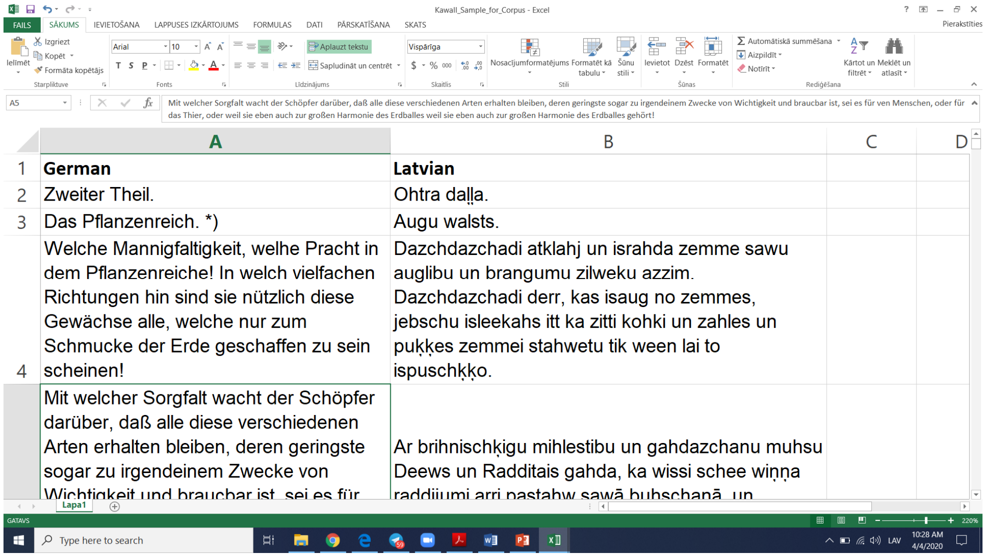
Feature Lists

- Gender
  - female
  - male
  - neutral
- Mood
  - imperative
  - indicative
- Number
  - dual
  - plural
  - singular
- Part of Speech
  - adjective
  - adverb
  - cardinalNumber
  - noun
  - ordinalNumber

Figure 2. Virtual Meeting with Dr. Karlheinz Mörth

Aligning the original work and its translation.

Aligning was performed by taking a sample from Chapter 2 of both books and manually copying the original work and translation segments into Excel Sheets (see Figure 3), as Excel is one of the formats required for future work with Sketch Engine. As the German text was relatively erroneous and required a lot of manual editing work, only a small part of the text could be processed in this way during the research visit.



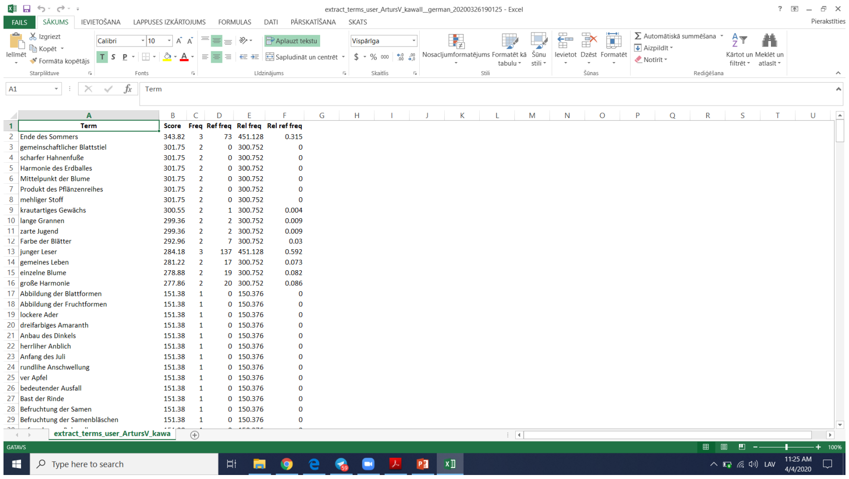
Mit welcher Sorgfalt wacht der Schöpfer darüber, daß alle diese verschiedenen Arten erhalten bleiben, deren geringste sogar zu irgendeinem Zwecke von Wichtigkeit und braucbar ist, sei es für ven Menschen, oder für das Thier, oder weil sie eben auch zur großen Harmonie des Erdballes weil sie eben auch zur großen Harmonie des Erdballes gehört!

	A	B	C	D
1	**German**	**Latvian**		
2	Zweiter Theil.	Ohtra daļļa.		
3	Das Pflanzenreich. *)	Augu walsts.		
4	Welche Mannigfaltigkeit, welhe Pracht in dem Pflanzenreiche! In welch vielfachen Richtungen hin sind sie nützlich diese Gewächse alle, welche nur zum Schmucke der Erde geschaffen zu sein scheinen!	Dazchdazchadi atklahj un israhda zemme sawu auglibu un brangumu zilweku azzim. Dazchdazchadi derr, kas isaug no zemmes, jebschu isleekahs itt ka zitti kohki un zahles un puķķes zemmei stahwetu tik ween lai to ispuschķķo.		
	Mit welcher Sorgfalt wacht der Schöpfer darüber, daß alle diese verschiedenen Arten erhalten bleiben, deren geringste sogar zu irgendeinem Zwecke von Wichtigkeit und brauchbar ist, sei es für	Ar brihnischķigu mihlestibu un gahdazchanu muhsu Deews un Radditais gahda, ka wissi schee wiņņa raddiiumi arri pastahw sawā buhschanā, un		

Figure 3. Aligned texts in Excel Sheets

Extracting data for a bilingual glossary.

To obtain data for the glossary, the Sketch Engine software was used. Texts in Excel tables were first uploaded there, and two text corpora were created from those – German and Latvian language corpus. Using the functionalities of Sketch Engine extractions, keyword lists and term lists with detailed information were created, e.g. score, frequency (see Figure 4 for German). They are intended for future use in creating glossaries and in terminology research.



Term	Score	Freq	Ref freq	Rel freq	Rel ref freq
Ende des Sommers	343.82	3	73	451.128	0.315
gemeinschaftlicher Blattstiel	301.75	2	0	300.752	0
scharfer Hahnenfuße	301.75	2	0	300.752	0
Harmonie des Erdballes	301.75	2	0	300.752	0
Mittelpunkt der Blume	301.75	2	0	300.752	0
Produkt des Pflänzenreihes	301.75	2	0	300.752	0
mehliger Stoff	301.75	2	0	300.752	0
krautartiges Gewächs	300.55	2	1	300.752	0.004
lange Grannen	299.36	2	2	300.752	0.009
zarte Jugend	299.36	2	2	300.752	0.009
Farbe der Blätter	292.96	2	7	300.752	0.03
junger Leser	284.18	3	137	451.128	0.592
gemeines Leben	281.22	2	17	300.752	0.073
einzelne Blume	278.88	2	19	300.752	0.082
große Harmonie	277.86	2	20	300.752	0.086
Abbildung der Blattformen	151.38	1	0	150.376	0
Abbildung der Fruchtformen	151.38	1	0	150.376	0
lockere Ader	151.38	1	0	150.376	0
dreifarbiges Amaranth	151.38	1	0	150.376	0
Anbau des Dinkels	151.38	1	0	150.376	0
herrliher Anblich	151.38	1	0	150.376	0
Anfang des Juli	151.38	1	0	150.376	0
rundlihe Anschwellung	151.38	1	0	150.376	0
ver Apfel	151.38	1	0	150.376	0
bedeutender Ausfall	151.38	1	0	150.376	0
Bast der Rinde	151.38	1	0	150.376	0
Befruchtung der Samen	151.38	1	0	150.376	0
Befruchtung der Samenbläschen	151.38	1	0	150.376	0

Figure 4. Extracted Terms from German corpus

Short conclusions and future prospects after research visit.

To sum up the experience, it must be concluded that the prepared text material plays a significant role. In this case the text was in old script in two languages (German and Latvian). The greatest challenge of this research was to successfully prepare the old script material, and it needs to be further developed. Another objective would be to research the available software for editing German Gothic script, as manual editing work is time-consuming. The result of the research visit are digitised versions of two sections of printed books mentioned above and extracted keyword and term lists for the text samples. The extracted material will be further analysed by frequency data, as well as clarifying the most popular word collocations. In this way it will be possible to identify specific translation models and strategies in terminology translation developed by H. Kawall. The extracted Latvian botanical terms used in H. Kawall's work will also be analysed in relation to the contemporary botanical terminology, and the results of the research in the form of an article will be included in the multilingual "New Botanical Dictionary" (a mobile app prototype), developed at Ventspils University of Applied Sciences.

Special thanks to: Dr. Tanja Wissik, Dr. Karlheinz Mörth and Mag. Anna Woldrich, as well as for the possibility of this remote stay during the ELEXIS visiting grant, taking into consideration the restrictions made due to the corona crisis.

elexis european lexicographic infrastructure

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

Universidad de Valladolid

ELEXIS Transnational Research Visit Grant (Call 3)

Final report

Grant holder: Lucía Sanz Valdivieso.

Hosting institution: Austrian Centre for Digital Humanities (Austrian Academy of Sciences).

Host: Tanja Wissik, PhD.

Institutional affiliation: English Philology Department, Facultad de Filosofía y Letras, University of Valladolid.

Project title: Description, creation and exploitation of online lexicographic and terminological resources for the Teaching of English for Specific Purposes.

Period of stay: December 2nd-6th, 2019.

1. Introduction

I applied for ELEXIS Translational Research Visit Grants because one of their host institutions called my attention: The Austrian Centre for Digital Humanities (ACDH) is involved in work focused on dissemination and interoperability, which are key factors in the development of new lexicographic tools. Besides, they also take part in research on training and education, which was also appealing since our project is aimed at being a pedagogical tool for learners of English for Specific Purposes. In sum, their institution's areas of work matched those of our project, and therefore the ACDH offered different methodological and technological possibilities to develop an interoperative lexicographic resource for the teaching of English for Specific Purposes.

2. Goals of the project

There is an increasing demand for the education of professionals-to-be in the competent use of English for Specific Purposes, which nowadays needs to be addressed from a digital and interoperability-oriented perspective. Because of this, the main goal of the project is the development of a pedagogical terminological tool targeted at Spanish native learners of English for Specific Purposes of their field, more particularly, the olive oil and wine tasting fields. This field and its LSP are highly relevant both at a regional and national levels, as Spain is the world leader in the exportation of these products, and most tourists who enjoy *oenotourism* and *oleotourism* are from foreign countries, thus needing to communicate in today's *lingua franca*, English.

In this context, our project is aimed at developing a Spanish-English/English-Spanish bilingual, multidimensional, and multifunctional dictionary which will be a complete pedagogical tool to provide full linguistic assistance by including not only terminological but also grammatical information (collocations, actants, syntactic behaviour). Besides, it will include encyclopaedic information and skill-oriented activities for learners to have a tool that covers both their communicative and pragmatic needs during their learning process. In this sense, the situations previewed to be involved in the use of the dictionary are production, reception and translation of texts both in their mother tongue (Spanish) and in the foreign language (English).

3. Description of the Research Visit

My research visit to the Austrian Centre for Digital Humanities started on December 2nd with an exciting schedule for the week. First, I attended a Research Lunch at the ACDH, where a lexicographic project that is being developed at the Centre was presented. There, I got to know about the updates on their dictionary of Bavarian dialects in Austria, the WBÖ (Wörterbuch der bairischen Mundarten in Österreich). This was especially enriching, as this is a project which started more than 100 years ago and therefore provides a remarkable illustration of how lexicographical methodological practice has evolved from manually-entered handwritten dictionary articles, to the currently-used XML/TEI Lex-0 format. I found it extremely useful for our project to learn about this format, which is open, easy to read, strictly structured, and, above all, allows for modification and enrichment of information and its interoperability.

Then, my host Tanja Wissik had scheduled a visit to the CLARIN Knowledge Centre for Terminology Resources and Translation Corpora (TRTC) at the University of Vienna. There, two experts on these fields, B. Heinisch and V. Lušicky, received us to have a really interesting discussion on terminology, translation, and standards. Among other things, they provided many valuable insights so as to how to embetter our project. Besides, they showed me some of the work they are currently involved in. Particularly, they mentioned the creation of UniVieTerm (recommendations and preferences for German and English terms within the university sphere). The discussion we had around it was extremely advantageous for our project, as they had implemented many features oriented at the interaction with users, something we had not considered initially. For example, the fact that users can *like* the articles and entries they find especially useful, or suggest new terms or corrections of existing terms in the database, offers an interesting option to be included in any pedagogical tool. We also discussed other initiatives aimed at the standardization of terminology, such as the Sprachressourcenportal Österreichs, a portal for resources involving administrative language and its cross-communication in German and English.

Next, I had the opportunity to share our project with peers at the ACDH. There, I explained an overview of the justification of the need of a pedagogical terminological tool on the fields of olive oil and wine tasting in Spain and gave examples of these particular Languages for Specific Purposes. I showed the material we are working with and how far we are into the project. In this sense, because we are at a very preliminary stage, all the discussion and questions that came after my presentation of the project were extremely useful and very likely to be implemented in the future outcome. Because these peers saw the project with a fresh perspective, some of the situations we had previewed for the users of the tool no longer seemed relevant: an instance is the translation from the foreign language (English) to the mother tongue (Spanish) we had foreseen as a situation

elexis
Description, creation and exploitation of online lexicographic and terminological resources for the teaching of English for Specific Purposes

in which learners would need assistance. However, we discussed about the fact that the tool stemmed from the need of internationalization of the products in question and of the competence of the professionals involved, and thus it makes much more sense to focus just into inverse translation from Spanish into English instead of trying to cover both translation directions. This is one example of the most interesting outcomes from this Research Lunch, although there were many questions regarding the project which altogether contributed either to support or discard the various features and options we had in mind for our dictionary based on its purpose.

Finally, I had the opportunity to have a meeting with Karlheinz Mörth, who is currently serving as the director of the Austrian Centre for Digital Humanities, to discuss tools and developments in lexicography. It was a fruitful meeting, as I was introduced to TEI Lex-0, a standard which is necessary nowadays for the adequate development of dictionaries. My host, Tanja Wissik, had also told me about it and shared with me some papers to read as an introduction preliminary to start learning encoding in this standard. K. Mörth showed me how this standard works by illustrating me through the Vienna Corpus of Arabic Varieties (VICAV). It is a project aimed at collecting digital language resources documenting varieties of spoken Arabic, and we went through it paying attention to article structure encoded in TEI Lex-0. We interchanged impressions on how differently lexicographers may work, and ended up agreeing that, as linguists and lexicographers, it is essential to have knowledge of how digital resources are built in order to have a better understanding of the research possibilities we can take advantage of for our projects. It was more than inspiring to listen to his wide experience as a lexicographer and discuss with him our project and some of his past and present projects in lexicography.

4. Concluding remarks

Our main goal in this Research Visit was to get to know how experts in lexicography develop their projects in institutions that use pioneering technology. In this sense, the Visit was a success, as I learnt more than I could have imagined when I submitted my project proposal for the ELEXIS Transnational Research Visits Grants. It has been successful both regarding the improvement of the dictionary design and the technical and methodological approaches we now plan to take in the actual development of the dictionary we intend to produce. Thanks to the many moments I have shared with experts and professionals in the field, our project will comply with standards that will allow us for easier modification of information and to ensure the interoperability of our data among systems but also across time. Plus, learning about the many projects that are being undertaken right now by these professionals has widened my perspective on my own project, which I believe has benefited enormously after this knowledge-sharing and feedback contact. All in all, this experience has contributed to my instruction to a large extent and has been hugely motivational for me as a researcher.

5. Acknowledgements

I want to thank ELEXIS for the opportunity they gave me to get to know the most up-to-date methods of working in lexicography. I would also like to thank Tanja Wissik for her wonderful efforts while being my host in the ACDH, and also every expert who took a bit of their time to share with me their knowledge and experience, and their impressions and suggestions regarding my project. And last, but not least, thanks to Belén López Arroyo (University of Valladolid) for trusting me to take our project to the ACDH for its discussion and enrichment.

ELEXIS 2019 Transnational Research Visit Report
Encoding definitions with word embeddings for sense categorization and diachronic linguistic studies

Luis Espinosa Anke
School of Computer Science and Informatics
Cardiff University, UK

1 Motivation

This document summarizes the work carried out by Luis Espinosa-Anke during his 2-week stay at the Royal Spanish Academy in the context of an ELEXIS transnational research travel grant. His work was focused on learning concept representations with distributional (i.e., corpus-based) and lexicographic (i.e., from a concept's definition) information. These concept representations, regardless of the approach, can be used in a number of downstream applications, from contributing to the dictionary writing and querying process, to improving the performance of NLP systems requiring semantic understanding, and even for enabling corpus-driven sociolinguistic analyses. At the end of the stay, three sub projects emerged, which will be continued during the following months, aiming at disseminating the most relevant findings in NLP and/or lexicography venues.

2 Tasks

2.1 Sense categorization

Semantic clustering can improve, on one hand, information access from a resource end user perspective [3]. Moreover, the acquisition of domain-specific training data for developing downstream language technologies has yielded promising results in NLP tasks such as word sense disambiguation [11, 1], text categorization [12] or hypernymy modeling [5]. In fact, in the context of ELEXIS, this idea has also been pursued, although from a slightly different angle, as we will see[1].

Given that there is literature in the NLP domain arguing for leveraging both corpus-based statistics and dictionary definitions to improve concept representations [9, 13, 2], we propose to explore the clustering of concepts in the Dictionary of the Spanish Language (*Diccionario de la Lengua Española* or *DLE*) based on semantic criteria, i.e., by determining the extent to which a definition can be mapped to a vector representation with and without the help of recurrent neural networks. We experiment with different approaches, which we discuss as follows.

Centroid-based clustering: As a simple baseline, we simply encode a concept by computing a weighted sum (by frequency) of the bag-of-words representation of its associated dictionary definition. This has the effect of grouping together, first, terms whose definitions share the *exact same words*, and after that, those where the different expressions are semantically similar to each other. In Table 1 we can see that while this approach seems desirable when aiming at grouping synonym terms with highly similar phraseology, it may also group terms which do not share any outstanding semantic feature, but happen to be defined using the same formulaic expressions.

Autoencoding average-based definitions: After observing the perhaps undesired behaviour of the semantic clusters emerging from the centroid-based strategy, and inspired by [6], we apply a conditional autoencoder to the average-based definition embeddings. Intuitively, we are interested in training a neural network capable to accurately reconstruct the original averaged vector, but by only using the information provided in content words (and not functional or stopwords, which we argue are more linked

[1] https://elex.is/wp-content/uploads/2019/08/ELEXIS_D3_1_Lexical_semantic_analytics_for_NLP_sense_clustering_Final.pdf

CENTROID-BASED	
cuerda	Extensión o número de notas que alcanza la voz
tenor	Hombre que tiene voz de tenor
atiplar	Dicho de la cuerda de un instrumento, o de la voz: Volverse del tono grave al agudo
AUTOENCODED CENTROID-BASED	
cuerda	Extensión o número de notas que alcanza la voz
subir	Dicho de la voz o del sonido de un instrumento: Pasar a un tono más agudo
sincopado/da	Dicho del ritmo o del canto: Que tiene notas sincopadas
LSTM-BASED	
atenorado/da	Dicho de una voz: Parecida a la del tenor
atenorado/da	Dicho de un instrumento: Que tiene un sonido de timbre semejante al de la voz del tenor
atiplar	Dicho de la cuerda de un instrumento, o de la voz: Volverse del tono grave al agudo

Table 1: Most similar terms (left column) and definitions (right column) to the target term "do de pecho" (*C major*), defined as "Una de las notas más agudas a que alcanza la voz de tenor" (*one of the highest tones that a tenor's voice reaches*).

to formulaic and stylistic patterns). For this reason, our autoencoder has, at decoding time, access to *the average vector of the stopwords of the definition*. In a way, this enforces the network to not consider any formulaic information when reconstructing the original vector from a compressed representation, as this information is explicitly provided. We can see in the comparison table that slightly different clusters emerge for the same target concepts (and definitions), with the interesting byproduct that these autoencoded representations can be tuned. In [6], it was found, in a different but related distributional semantic task, that going as low as 10-dimensional vector provided a fine balance between abstraction and interpretability. The dimensionality of the definition vectors provided in Table 1 is 25 (empirically set based on qualitative analysis).

LSTM-based clustering: Recent literature in computational lexicography and computational semantics has argued for taking advantage of the mapping power of recurrent neural networks to produce a vector representation of a definition such that it resembles as much as possible the vector representation of the definiens (i.e., the term being defined). This is useful because one single neural network architecture can be trained over a full dictionary resource, and from such model the hidden representation of the last LSTM [10] state can be thought of encoding the semantics of the definition, as well as the mapping history from previous ⟨term, definition⟩ pairs. As we can see in the selected examples in Table 1, this approach is better at downweighting formulaic (and semantically void) definitional expressions, while at the same time retaining the most salient semantic features of each definition.

2.2 Thematic category classification

The DLE provides a manually constructed thematic classification, which can be thought of a grouping of dictionary entries by their belonging to domains of knowledge[2]. The Spanish Academy, on the other hand, currently holds a vast number of unannotated resources, ranging from diachronic to to domain-specific (e.g., legal) dictionaries. As a proof of concept, we trained a CBLSTM classifier (convolutional layer followed by a bidirectional LSTM), which has proven effective in previous work for definition modeling [7]. We trained it on the coarsest-grained categorization of thematic labels on the DLE (which contains 8 classes), and generated a thematically-tagged Wiktionary-ES version. Because we have used a set of cross-lingual word embeddings [4], such thematic categorization could easily be ported to other languages. This constitutes an interesting case of transfer learning, where English can be considered the target language, and Spanish, due to the availability of the Academy's resources, the (resource-rich)

[2] See a user guide at `https://enclave.rae.es/pdfs/Guia_de_uso\%20Enclave_RAE.pdf`.

source language. To get an idea of the difficulty of the task, we show below in Figure 1 a PCA projection of the centroid-based representation of the DLE definitions and their different thematic categories. We can see, for example, that definitions categorized as 'el mundo' (*the world*) are easier to classify from the rest, whereas 'vida humana' (*human life*) and 'ciencias humanas' (*human science*) are almost undistinguishable from each other.

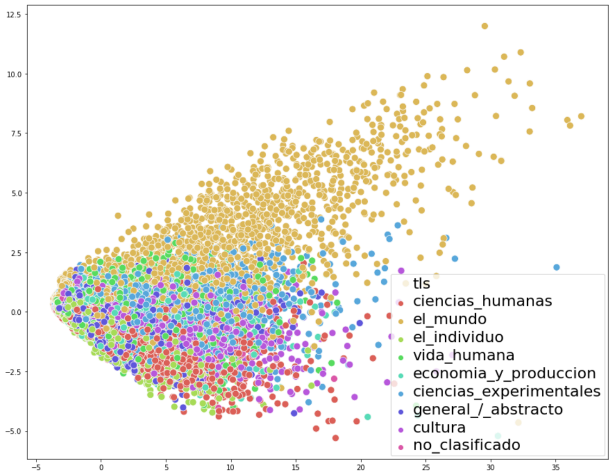


Figure 1: PCA visualization of DLE definitions and their grouping according to their thematic categorization.

2.3 Semantic Change

An interesting application of vector representations of words is that they can be used to assess the extent to which a word's meaning changes (or drifts), as this change is correlated with different contexts in which the word is used, and therefore, different co-occurrence statistics will emerge, yielding a different position in the vector space. More importantly, the nearest neighbours of the word vector at different timeframes will be different, as these are a reflection of its meaning. This intuition was explored using corpora of different centuries in [8]. During our ELEXIS project, and due to the availability of diachronic corpora from the 1970s of the largest newspaper in the Spanish language (El País[3]), we started working on a piece of software for exploring the semantic drift of target words in the Spanish language, and decided empirically to do this over a 5-year window from the earliest to the latest available documents (effectively, from 1977 to 2011). An illustrative example of some words which have clearly changed meaning in these three decades are shown in Table 2. We see that for 'rescate', the criminal sense (a rescue of hostages) was more predominant decades ago, while in its modern form is more associated with financial rescuing or bailout. Another interesting example is the word 'entertainment' (*entretenimiento*), which in the past was more associated to TV, and now it is more linked to digital entertainment (with similar words like 'interactivo' (*interactive*) and 'online'.

2.4 Conclusions and future work

This research stay has had three major impacts:

- **Strong collaboration**. A strong line of collaboration between the Academy and Cardiff University has been established, and we plan to turn it into one or more submissions to NLP or Lexicography forums. In fact, the Academy is currently considering bringing into their regular pipeline (both for lexicographers and for user queries) access to semantic clusters emerging from word embeddings and neural network-based encodings.

- **Insights on the Spanish language**. We have learned a lot about how the Spanish language works,

[3] `www.elpais.es`.

RESCATE		**CARTERA**		**ENTRETENIMIENTO**	
77-81	*07-11*	*77-81*	*07-11*	*77-81*	*07-11*
salvamento	salvamento	pedidos	carteras	instrumento	interactivo
operación	Grecia	carteras	diversificación	programas	online
secuestradores	emergencia	Industrias	inversión	vídeo	televisivo
pagado	pecuniario	Agricultura	activos	ocio	diversión
reparación	rescates	Finanzas	pedidos	visual	ocio

Table 2: Nearest neighbours for three target words: 'rescate' (*rescue* or *bailout*), 'cartera' (*wallet* or *portfolio*) and 'entretenimiento' (*entertainment*).

both in terms of how its words are defined in its reference dictionary, but also how they are classified and how their meaning has changed over time.

- **Researcher growth**. Personally, I am very satisfied by the opportunities for development that this grant has provided. I am confident that I am now a better researcher and that I know more about language and lexicography than I did before the stay, and I am looking forward to a future of collaborations.

For the future, in addition to turning these initiated projects into tangible submissions, we plan to extend our work to the cross-lingual setting, and to focus on objective and measurable success criteria for the clustering algorithms. Finally, we would like to incorporate a dimension of *relational similarity*, which due to lack of time and unclear straightforward application, was not tackled during the stay.

References

[1] Eneko Agirre, Oier Lopez De Lacalle, and Aitor Soroa. Knowledge-based wsd on specific domains: performing better than generic supervised wsd. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[2] Tom Bosc and Pascal Vincent. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, 2018.

[3] Jose Camacho-Collados and Roberto Navigli. Babeldomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228, 2017.

[4] Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. Improving cross-lingual word embeddings by meeting in the middle. *arXiv preprint arXiv:1808.08780*, 2018.

[5] Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. Extasem! extending, taxonomizing and semantifying domain terminologies. In *AAAI*, 2016.

[6] Luis Espinosa-Anke and Steven Schockaert. Seven: Augmenting word embeddings with unsupervised relation vectors. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2653–2665, 2018.

[7] Luis Espinosa-Anke and Steven Schockaert. Syntactically aware neural architectures for definition extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 378–385, 2018.

[8] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

[9] Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30, 2016.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Bernardo Magnini and Gabriela Cavaglia. Integrating subject field codes into wordnet. In *LREC*, pages 1413–1418, 2000.

[12] Roberto Navigli, Stefano Faralli, Aitor Soroa, Oier de Lacalle, and Eneko Agirre. Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2317–2320. ACM, 2011.

[13] Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec: Learning word embeddings using lexical dictionaries. 2017.

Exploring Digitization and Encoding Options for Ben Yehuda’s Hebrew Dictionary

Sinai Rusinek: report on a week-long STRV visit to BCDH and ELEXIS-RS.

The seminal “Dictionary of Old and New Hebrew” was established as a significant step in the revival of Hebrew as a literary language, which started in Eastern Europe in the 18th century, and culminated in Eliezer Ben Yehuda’s life’s project, to make Hebrew into a spoken, living language. The publication of the dictionary lasted more than five decades and outlived Ben Yehuda, who in his lifetime saw the publication of only 5 of the 17 volumes.

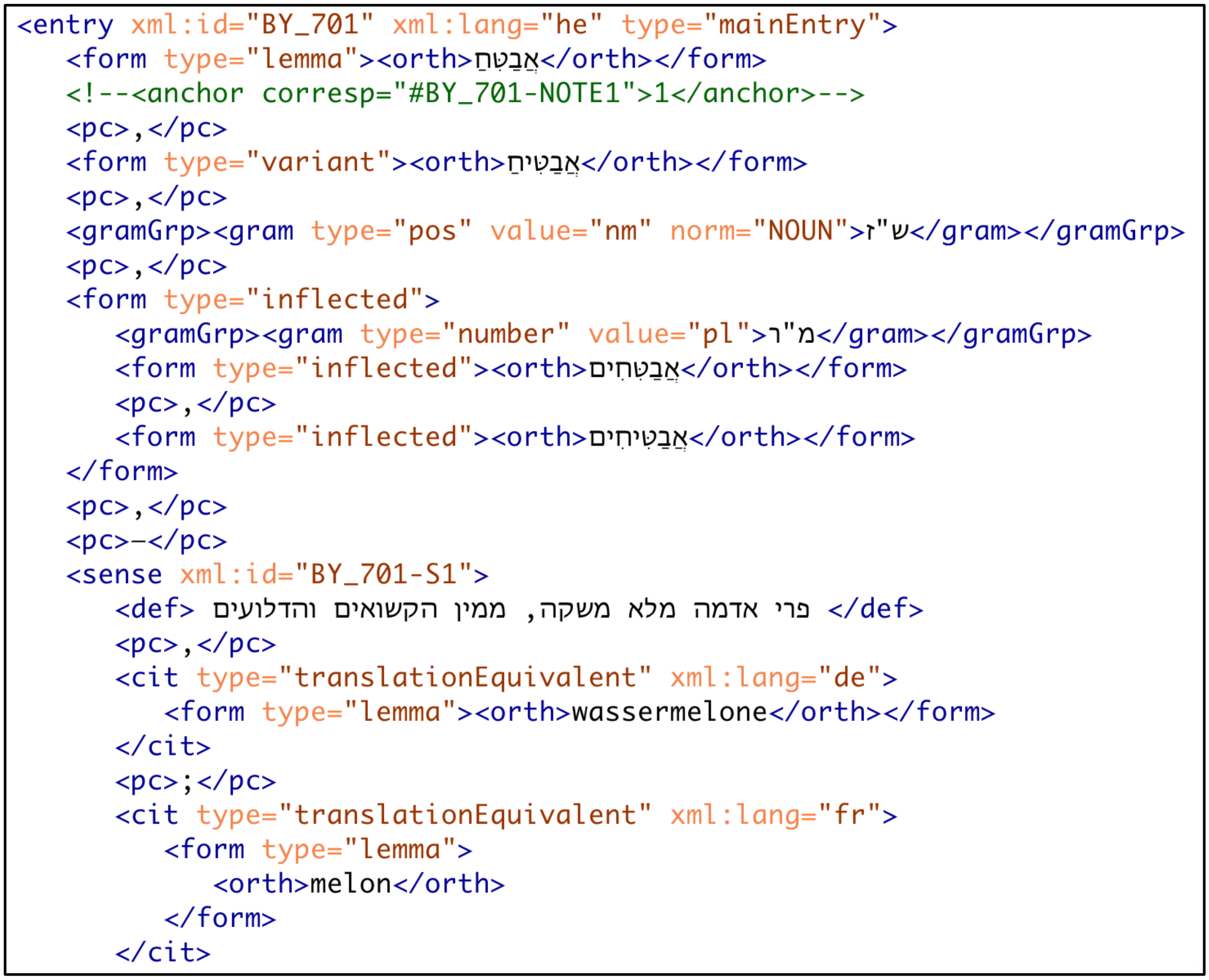
The dictionary contains almost 8000 words and the entries include translations to German, French and English, Hebrew definitions, quotations from historical and contemporaneous sources and often etymological and other notes, connecting old words as well as neologisms to both Asian and European traditions, as well as to ideas about science, technology and progress. This plethora make the dictionary into an invaluable resource for the study of historical semantics, ethnography and cultural history of the period. Its main potential, however, can be fully realized when it has undergone a refined digitization process.

The digitization project of Ben Yehuda’s dictionary is a crowdsourcing volunteer based project which is taking place in the framework of the larger Ben Yehuda Project, the “Israeli Gutenberg project” https://bybe.benyehuda.org/page/english, founded and spearheaded by Asaf Bartov of the Wikimedia foundation, who has also built the transliteration environment for the dictionary project. 2400 entries have already been transcribed. The result of the transcription, which is done manually, is a partly-structured HTML. The web interface enables basic search, and biblical quotations were already linked to the online Wikitext Hebrew Bible.

Funded by Elexis, a short term research visit in Elexis-RS at the Belgrade Center for Digital Humanities, hosted by Dr. Toma Tasovac, was dedicated to explore the potential of making the Ben Yehuda dictionary a structured research resource, according to the TEI Lex-0 modelling convention.

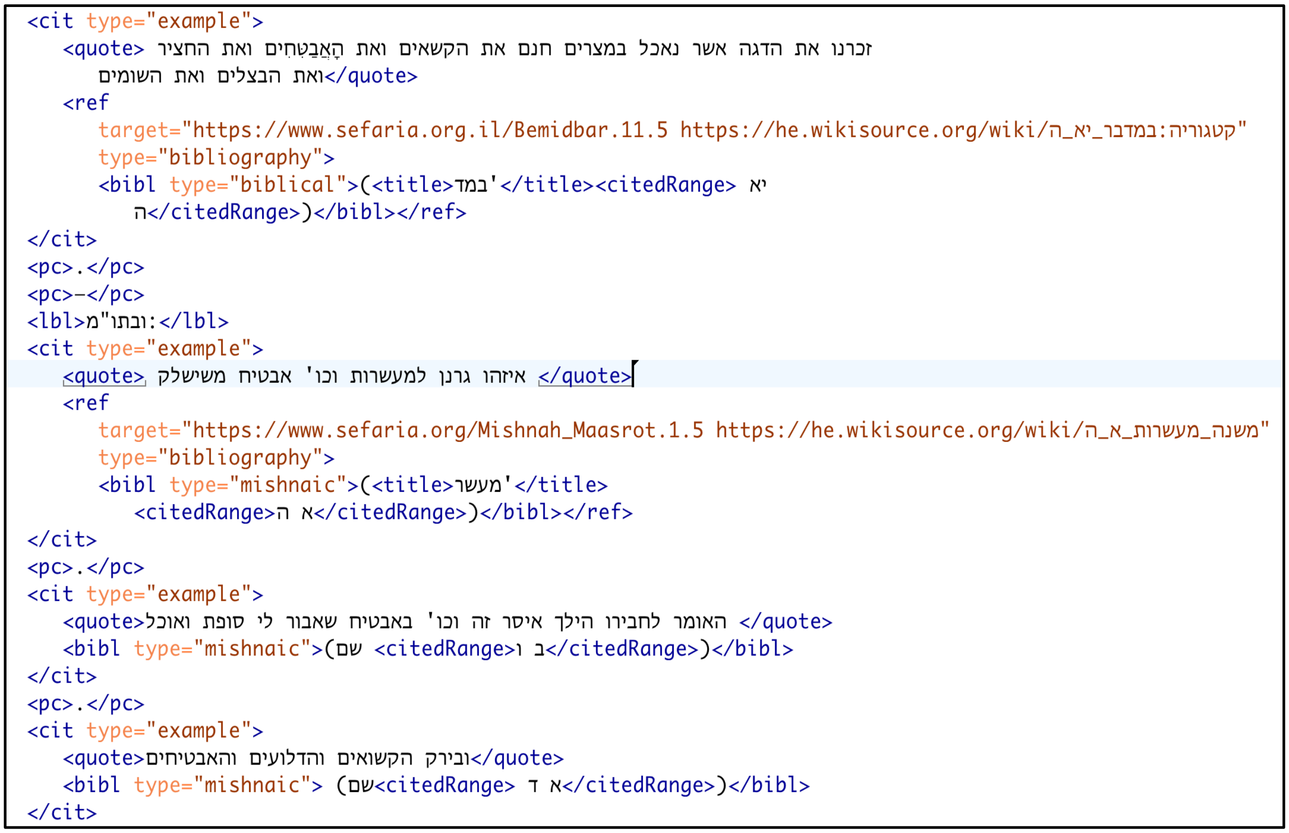
We discussed and experimented with platforms for publishing XML such as CETEICEAN and eXist-DB, which is the platform serving Raskovnik - the BCDH-designed platform for Serbian dictionaries. I was introduced to the oXygen and Github-based workflows for annotation and editing. We spent considerable time analyzing the microstructure of the dictionary, and, finally, we created a proof-of-concept annotation of sample entries from the Ben Yehuda dictionary. The modelling process shed light on the potential as well as the challenges the lie ahead for a future annotation project.

The Ben Yehuda dictionary is highly structured with syntactic information and rich in examples, citations and notes. The following example is a snippet from the encoding of a sample noun entry exhibiting these elements:



```
<entry xml:id="BY_701" xml:lang="he" type="mainEntry">
   <form type="lemma"><orth>אֲבַטִּחַ</orth></form>
   <!--<anchor corresp="#BY_701-NOTE1">1</anchor>-->
   <pc>,</pc>
   <form type="variant"><orth>אֲבַטִּיחַ</orth></form>
   <pc>,</pc>
   <gramGrp><gram type="pos" value="nm" norm="NOUN">ש"ז</gram></gramGrp>
   <pc>,</pc>
   <form type="inflected">
      <gramGrp><gram type="number" value="pl">מ"ר</gram></gramGrp>
      <form type="inflected"><orth>אֲבַטִּחִים</orth></form>
      <pc>,</pc>
      <form type="inflected"><orth>אֲבַטִּיחִים</orth></form>
   </form>
   <pc>,</pc>
   <pc>–</pc>
   <sense xml:id="BY_701-S1">
      <def> פרי אדמה מלא משקה, ממין הקשואים והדלועים </def>
      <pc>,</pc>
      <cit type="translationEquivalent" xml:lang="de">
         <form type="lemma"><orth>wassermelone</orth></form>
      </cit>
      <pc>;</pc>
      <cit type="translationEquivalent" xml:lang="fr">
         <form type="lemma">
            <orth>melon</orth>
         </form>
      </cit>
```

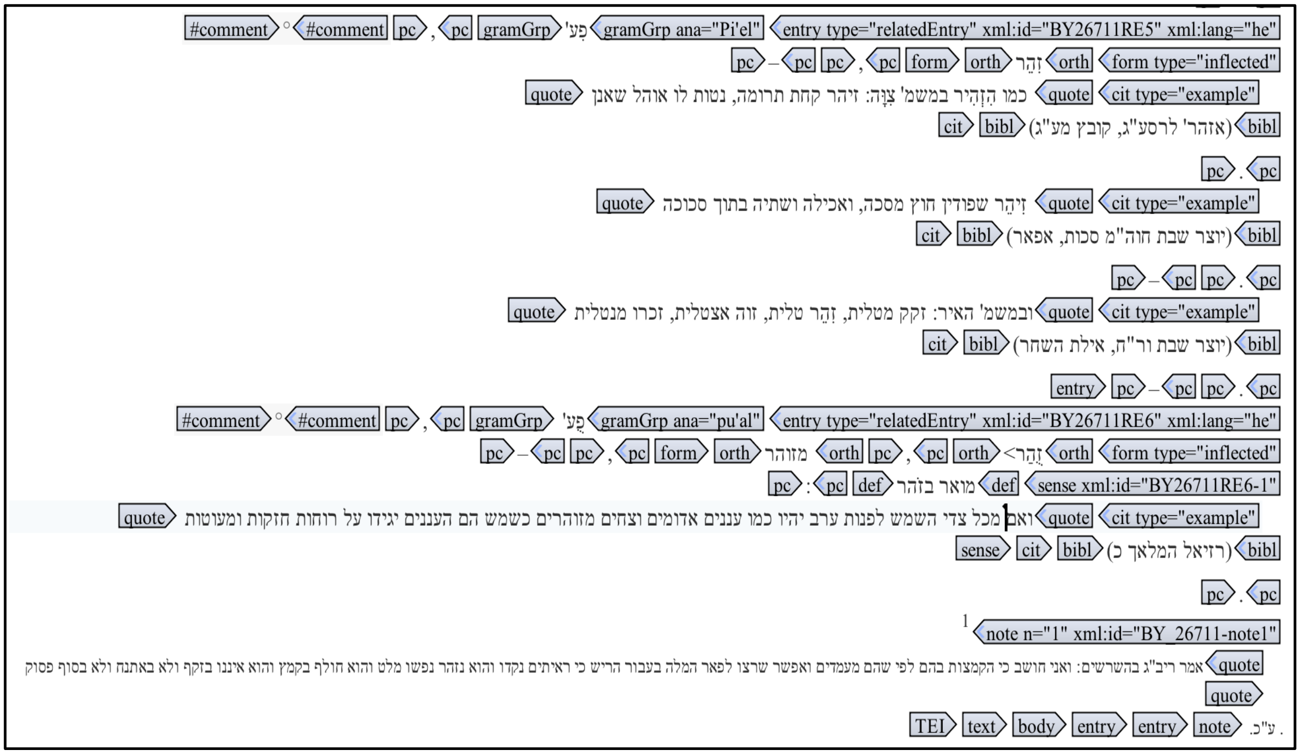
The citations, taken from biblical literature as well as from Mishnaic and Talmudic later sources, and to a lesser extent from modern Hebrew, is especially promising with regard to citation network analysis. A typology of the citation sources is required in addition to linking quotations with open resources where they may be read in context.



```
<cit type="example">
   <quote> זכרנו את הדגה אשר נאכל במצרים חנם את הקשאים ואת הָאֲבַטִּחִים ואת החציר
      ואת הבצלים ואת השומים</quote>
   <ref
      target="https://www.sefaria.org.il/Bemidbar.11.5 https://he.wikisource.org/wiki/קטגוריה:במדבר_יא_ה"
      type="bibliography">
      <bibl type="biblical">(<title>במד'</title><citedRange> יא
         ה</citedRange>)</bibl></ref>
</cit>
<pc>.</pc>
<pc>–</pc>
<lbl>ובתו"מ:</lbl>
<cit type="example">
   <quote> איזהו גרנן למעשרות וכו' אבטיח משישלק </quote>
   <ref
      target="https://www.sefaria.org/Mishnah_Maasrot.1.5 https://he.wikisource.org/wiki/משנה_מעשרות_א_ה"
      type="bibliography">
      <bibl type="mishnaic">(<title>מעשר'</title>
         <citedRange>א ה</citedRange>)</bibl></ref>
</cit>
<pc>.</pc>
<cit type="example">
   <quote>האומר לחבירו הילך איסר זה וכו' באבטיח שאבור לי סופת ואוכל </quote>
   <bibl type="mishnaic">(שם <citedRange>ב ו</citedRange>)</bibl>
</cit>
<pc>.</pc>
<cit type="example">
   <quote>ובירק הקשואים והדלועים והאבטיחים</quote>
   <bibl type="mishnaic"> (שם<citedRange> א ד</citedRange>)</bibl>
</cit>
```

There are, however, some specificities of the dictionary and language that defy simple annotation: for example, inflected forms may appear in the dictionary as properties of a specific entry, or as different entries in a group. A sense is often not explicated for different inflections or form, as it is assumed to be obvious from the Hebrew inflection system. In these points the legacy dictionary conflicts with the expected regularity of the schema.

Also, an abundance of related entries is creating a complex hierarchy that calls for an elaborate system of unique IDs, covering the various categories (word family entry vs. main entry, nested (related) entry, and sense).



```
entry type="relatedEntry" xml:id="BY26711RE5" xml:lang="he" gramGrp ana="Pi'el" פִּע' gramGrp pc , pc #comment ° #comment
form type="inflected" orth זְהֵר orth form pc , pc pc – pc
cit type="example" quote כמו הִזְהִיר במשמ' צִוָּה: זיהר קחת תרומה, נטות לו אוהל שאנן quote
bibl (אזהר' לרסע"ג, קובץ מע"ג) bibl cit
pc . pc
cit type="example" quote זִיהֵר שפודין חוץ מסכה, ואכילה ושתיה בתוך סכוכה quote
bibl (יוצר שבת חוה"מ סכות, אפאר) bibl cit
pc . pc pc – pc
cit type="example" quote ובמשמ' האיר: זקק מטלית, זִהֵר טלית, זוה אצטלית, זכרו מנטלית quote
bibl (יוצר שבת ור"ח, אילת השחר) bibl cit
pc . pc pc – pc entry
entry type="relatedEntry" xml:id="BY26711RE6" xml:lang="he" gramGrp ana="pu'al" פֻּע' gramGrp pc , pc #comment ° #comment
form type="inflected" orth זֻהַר> orth pc , pc orth מזוהר orth form pc , pc pc – pc
sense xml:id="BY26711RE6-1" def מואר בזֹהר def pc : pc
cit type="example" quote ואם מכל צדי השמש לפנות ערב יהיו כמו עננים אדומים וצחים מזוהרים כשמש הם העננים יגידו על רוחות חזקות ומעוטות quote
bibl (רזיאל המלאך כ) bibl cit sense
pc . pc
note n="1" xml:id="BY_26711-note1" 1
quote אמר ריב"ג בהשרשים: ואני חושב כי הקמצות בהם לפי שהם מעמדים ואפשר שרצו לפאר המלה בעבור הריש כי ראיתים נקדו והוא נזהר נפשו מלט והוא חולף בקמץ והוא איננו בזקף ולא באתנח ולא בסוף פסוק
quote
ע"כ. note entry entry body text TEI
```

With an eye to automating as much of the process as possible, we formulated some preliminary rules based on a combination of regular expressions and XPath as a basis for a conversion scripts that will hopefully expedite the work in a future implementation of the project. I also familiarized myself with XProc-based sets of XSLT transformations which are used at BCDH for automatic and semi-automatic conversions of paper-based dictionaries to TEI.

I am grateful to BCDH and ELEXIS for giving me the opportunity to expand my knowledge and hone my technical skills.