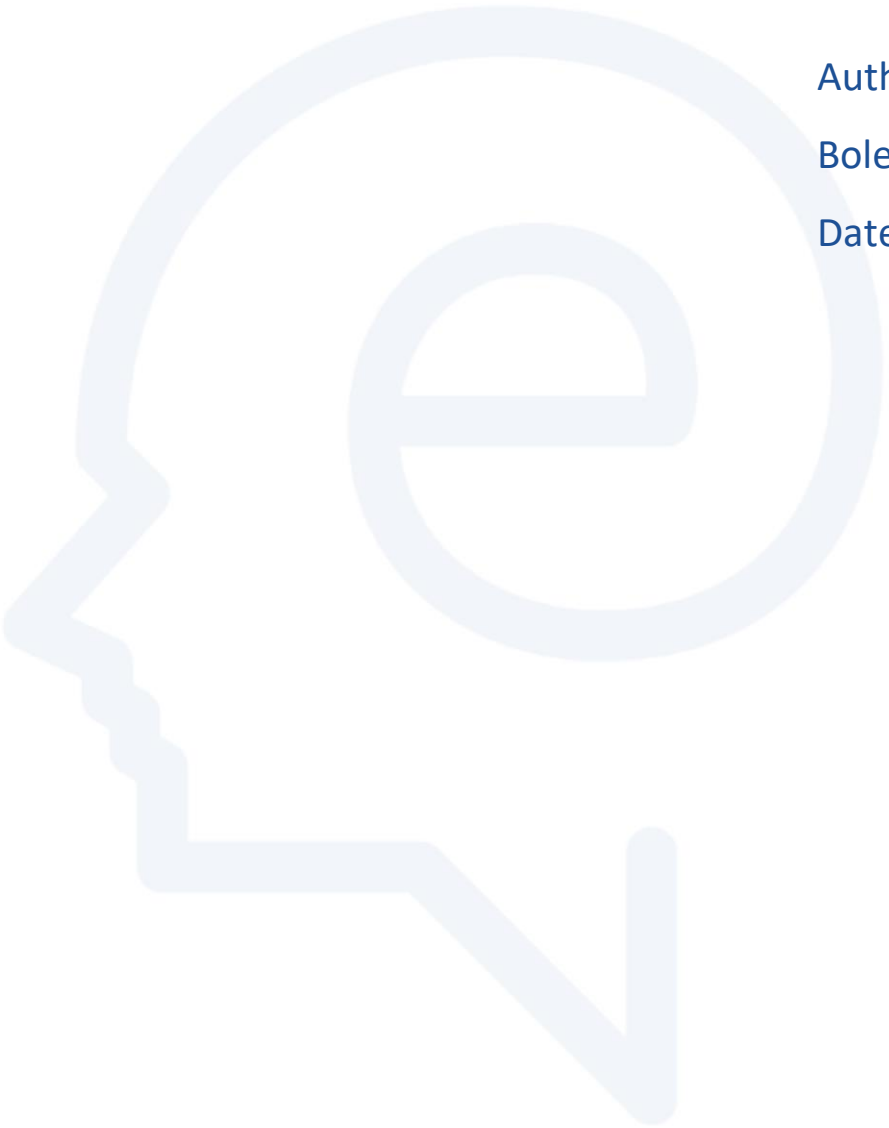
elexis european lexicographic infrastructure

D.9.4 Final report on trans-national access

Author(s): Sussi Olsen

Bolette S. Pedersen

Date: 31-07-2022



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. 9.4 Final report on trans-national access

Deliverable Number:	9.4
Dissemination Level:	PU
Delivery Date:	31-07-2022
Version:	V0.1
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 11-07-2022	First draft	Sussi Olsen, Bolette S. Pedersen
V02 25-07-2022	First draft review Second draft review Final version	Simon Krek Sussi Olsen, Bolette S. Pedersen

Final report on trans-national access

Table of Contents

1 Introduction: Trans-national access – year 4 and overall results

This deliverable presents the results of the final year of the transnational access program of the ELEXIS project as well as the accumulated results of the entire project period.

Due to the pandemic, only very few visits took place from spring 2019 to autumn 2021 so during spring and early summer 2022 several pending visits from previous calls were carried out. However, not all of the granted visits could be realised either due to ongoing restrictions or due to fact that people's projects had been finalised or their employment situation had changed meaning that a visit was no longer realistic or meaningful.

In the following sections, we present the results of the final call together with the accumulated results. We describe the challenges of the transnational activities caused by the Covid-19 crisis. Furthermore, we examine the profiles of all the grant holders, i.e. origin countries, gender and experience, and give an overview of the scientific output of the visits.

The objectives of the ELEXIS trans-national activities are summarised below:

- to offer opportunities to researchers or research teams to access research facilities with an excellent combination of advanced technology and expertise
- to support training of new specialists in the field of e-lexicography in order to conduct high-quality research and ensure sustainability of the infrastructure
- to ensure support for excellent scholarly research projects and innovative enterprises and also support the complex multi-disciplinary research
- to encourage the integrative use of technology and methodologies as developed in ELEXIS and in the lexicographical institutions
- to improve the overall services (lexicographic and technical) available to the research community
- to exchange knowledge and experience and to work towards future common projects and objectives
- to create an interdisciplinary community, collaborating on activities that are fully or partially of relevance to the proposed work of the grant holder
- to create knowledge at the interaction between academia and society

2 Call and grant winners, final call and in total

After postponing or cancelling calls during the pandemic and the consecutive lock down, the project launched a final and fifth call in December 2022. This was due to that fact that a couple of partners still found it possible to host visitors during spring/summer 2022.

The final call received seven applications of which four were rewarded with a grant. The four grant holders came from Croatia, Germany, Hungary, and Slovenia. Below is an overview of the home institutions, hosting institutions and project titles of these four grant holders.

Home institution	Hosting institution	Project
Old Church Slavonic Institute, Croatia	Institute for Bulgarian Language (IBL, Bulgaria)	Possibilities of digitalization and retrodigitalization and modernization of the dictionary writing process in *Dictionary of the Croatian Redaction of Church Slavonic*
Leibniz-Institut für Deutsche Sprache	Institute for Dutch Language (INT, The Netherlands)	How is the corpus influencing collocation sets in dictionaries?
Hungarian Office for Translation and Attestation Ltd. (OFFI Ltd.); Budapest University of Technology and Economics	Institute for Dutch Language (INT, The Netherlands)	The Description of the Characteristics of Legal Terminology and Issues of Editing Legal Databases in Lexicography and Terminology
Digital Signal Processing Laboratory, Faculty of Electrical Engineering and Computer Science, University of Maribor	Institute for Bulgarian Language (IBL, Bulgaria)	Metaphorical expressions in written and spoken language: from metaphor identification to metaphor detection

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

Overall, 22 visits were completed during the entire project period while five were cancelled[1]. Table 2 gives an overview of all the visits completed, showing the grant holders home institution, the hosting infrastructures and the title of the projects.

Home institution	Hosting institution	Project
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

[1] Note the one visit was completed just before writing this report (visit to JSI); all tables have therefore not been adjusted correspondingly, and the final report from this visit is still in progress.

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
Haifa University	Belgrade Center for Digital Humanities (BCDH, Serbia)	Exploring Digitization and Encoding Options for Ben Yehuda's Hebrew
Austrian Centre for Digital Humanities (Austrian Academy of Sciences), Austria	Real Academia Española (RAE, Spain)	A map based data visualization application to analyze variants of the Spanish language
Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa/Centro de Linguística da Universidade NOVA de Lisboa, Portugal	Real Academia Española (RAE, Spain)	Introduction to the macro and micro structure of RAE: Diccionario de la lengua española for the work on the new Dicionário da Língua Portuguesa (DLP)
Linguistics Research Institute of the Croatian Academy of Sciences, Croatia	Trier Center for Digital Humanities (TCDH, Germany)	(Retro)digitisation and online publication of the Croatian Dictionary of the Literary Language
NOVA CLUNL - Universidade NOVA de Lisboa, Portugal	Dutch Language Institute (INT, The Netherlands)	A multisemiotic e-dictionary/Knowledge Organization and Terminology: application to Cork
Copenhagen Business School, Denmark	K Dictionaries (KD, Israel)	Business Model Innovation in the Dictionary Industry

Sofia University "St. Kliment Ohridski“, Bulgaria	Belgrade Center for Digital Humanities (BCDH, Serbia)	Encoding Latin-Bulgarian Dictionary
CELGA-ILTEC, University of Coimbra, Portugal	Institut Jožef Stefan (JSI, Slovenia)	Improving a procedure for automatic extraction of data and import into DWS
Institute of Croatian Language and Linguistics, Croatia	Det Danske Sprog- og Litteraturselskab (DSL, Denmark) & University of Copenhagen (UCPH, Denmark)	Nordic E-dictionaries in Comparison to the Croatian Web Dictionary – Mrežnik
Batumi State Maritime Academy, Georgia	Dutch Language Institute (INT, The Netherlands)	English - Georgian Maritime Dictionary
Old Church Slavonic Institute, Croatia	Institute for Bulgarian Language (IBL, Bulgaria)	Possibilities of digitalization and retrodigitalization and modernization of the dictionary writing process in Dictionary of the Croatian Redaction of Church Slavonic
Leibniz-Institut für Deutsche Sprache	Institute for Dutch Language (INT, The Netherlands)	How is the corpus influencing collocation sets in dictionaries?
Hungarian Office for Translation and Attestation Ltd. (OFFI Ltd.); Budapest University of Technology and Economics	Institute for Dutch Language (INT, The Netherlands)	The Description of the Characteristics of Legal Terminology and Issues of Editing Legal Databases in Lexicography and Terminology
Digital Signal Processing Laboratory, Faculty of Electrical Engineering and Computer Science, University of Maribor	Institute for Bulgarian Language (IBL, Bulgaria)	Metaphorical expressions in written and spoken language: from metaphor identification to metaphor detection
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia	Institut Jožef Stefan (JSI, Slovenia)	Adapting dictionary writing systems and other platforms to online dictionaries of idioms

Table 2. Overview of all visits showing grant holder affiliations, hosting institutions and project titles.

A scarcity of hosting institutions was foreseen after the first calls since the applications of the first three-four calls concentrated on a few popular hosting institutions who then almost spent

their visiting budget. However, due to the pandemic and consecutively fewer calls, this never became a real problem. Still, only two hosting institutions were open for the final call, as other hosting infrastructures either had pending postponed visits or still had national restrictions that made visits unmanageable.

In total, five visits that had been granted were cancelled. Figure 1 shows the total number of completed visits at the various hosting institutions together with the number of visits cancelled.

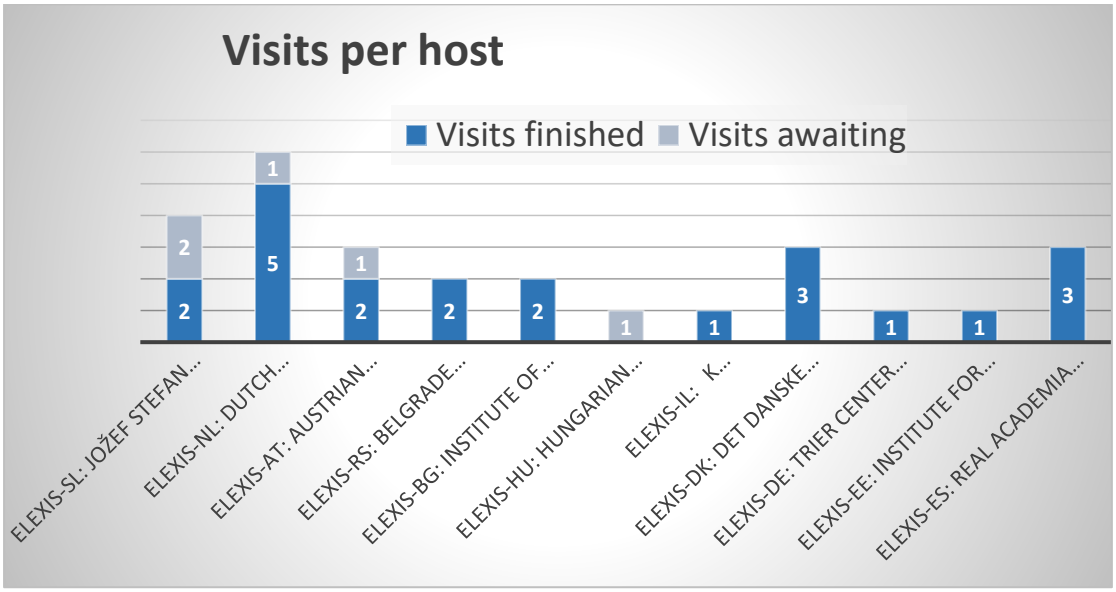


Figure 1. An overview of completed and cancelled visits per hosting institution.

3 The COVID-19 situation

The COVID-19 pandemic has not surprisingly been a huge obstacle to the ELEXIS transnational activities leading to a much lower number of activities than expected.

The foreseen number of visits throughout the project were 35-40 but due to the pandemic, all new calls were cancelled for almost two years. In total five calls were launched instead of the expected 7- 8 calls during the project, resulting in 22 visits carried out. Many granted visits have to be postponed several times and as already mentioned some of these ended up being cancelled.

The reason for the five cancellation were e.g. that the grant holder had left the field, was working on another project or it was simply not possible to find a time during spring 2022 that suited both grant holder and hosting institution.

The pandemic situation thus made it almost impossible for the hosting infrastructures to achieve the planned amount of visits (three per institution) and accordingly to spend the budget allocated to this task. It was therefore decided on a Project Management Board meeting in autumn 2021 that each partner should consider whether they would be able to spend the budget as planned or whether they should plan a re-allocation of the funds. This decision taken by individual partners is reflected in their final financial reports.

4 Profiles of grant holders

After the fourth call, we made a profile of the grant holders. This profile has been updated, including data from the final call to give a clear picture of who are the winners of the grants.

The grant holders come from a great variety of countries. The transnational program of ELEXIS clearly managed to reach communities all over Europe – and beyond. We received applications also from countries that are not EU or associated countries, e.g. Russia, China, Brazil, and Iran, but these could not be granted a visit. A list of the countries of the grant holders is shown in figure 2.

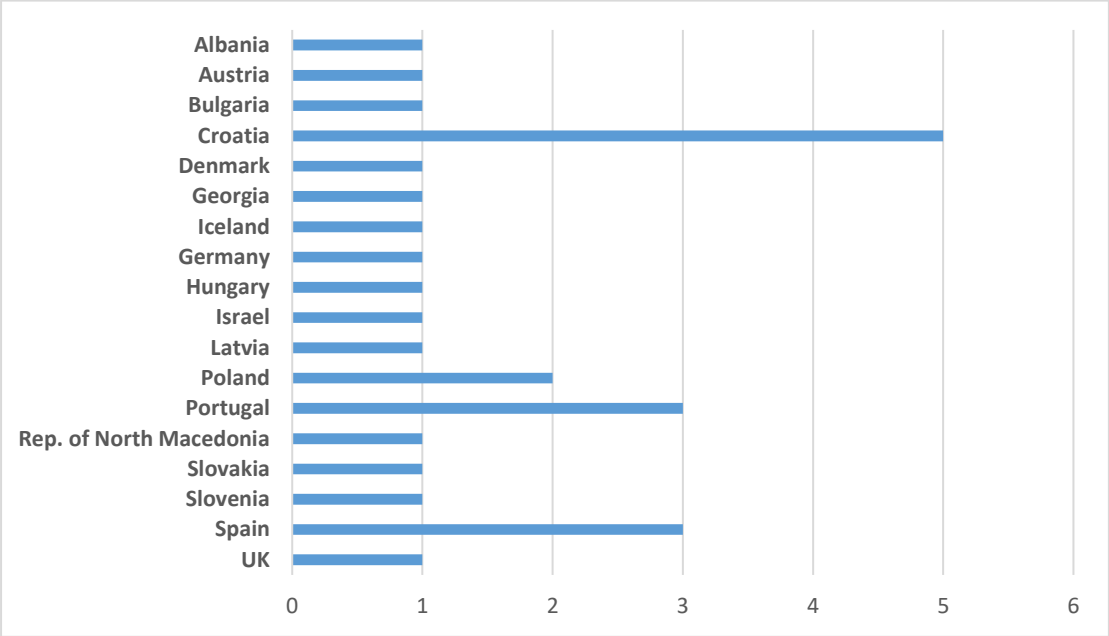


Figure 2. Countries of grant holders

We examined the gender of the applicants in total and compared it to those who have been awarded a grant. As can be seen in Figure 3, we received by far more applications from women than from men. In addition, the success rate for women is higher than for men.

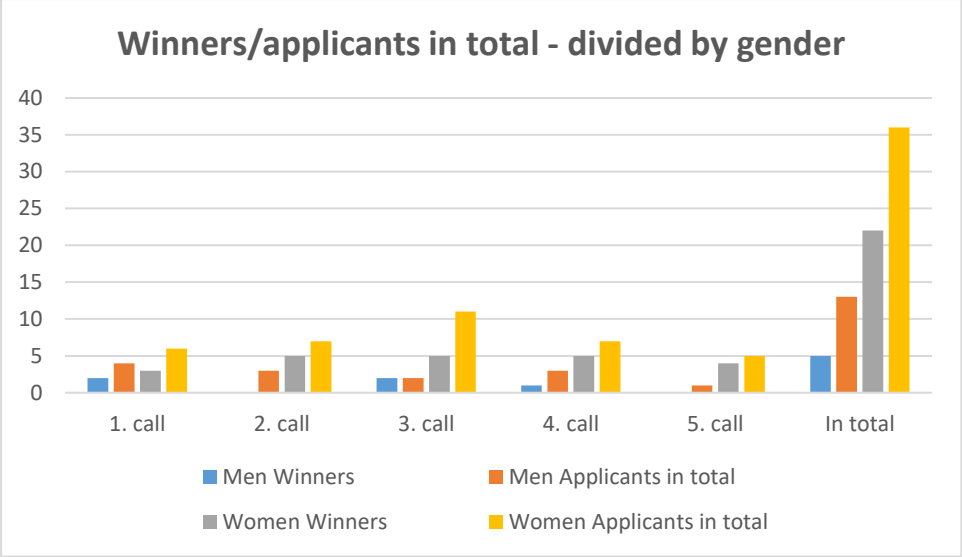


Figure 3: Applicants and winners – gender distribution for all calls

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.

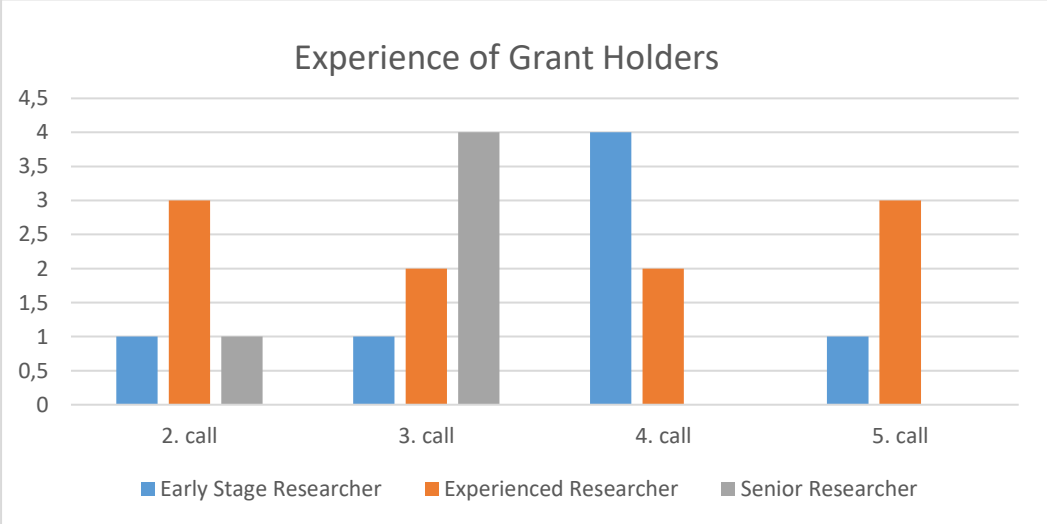


Figure 3: Experience of the grant holders

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 outcome of the completed visits

The grant holders (and hosts) from the final call in accordance with the other calls reported that the visits had been very rewarding 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 value of the networking and more general knowledge exchange is something that all grant holders report as a very important part of the visit.

Some typical topics of the grant holder projects:

- retro-digitalization of older/historical dictionaries
- dictionaries with terminological content
- ethical dilemmas in dictionary writing
- introduction to and comparison of dictionary structures and methodologies
- automatic extraction of corpus data for import into dictionaries

More specialised topics:

- map-based data visualization application
- business models of lexicography

The reports from the final call as well as from the visits from earlier calls finished the last year can be found in Appendix A.

All reports can be found on the ELEXIS website: https://elex.is/travel-grant-reports/.

The scientific outcome is further justified and disseminated through conferences where former visitors give talks about their projects, through PhD projects, which have been based on the project of the ELEXIS grant visit, and in scientific papers written by former grant holders. We see this as a clear indication that the transnational activities have satisfied an existing need.

6 Conclusion

Overall, the transnational activities of ELEXIS have been a great success. All grant holders report on high scientific, technical and networking value of these visits.

However, the activities could have reached out to so many more lexicographers and researchers and supported so many more lexicographic projects, had it not been for the almost two years of the COVID-19 pandemic. Almost twice as many visits were planned to have taken place and looking at the high number of applications for the final call in December 2021 where it was still unsure whether it would be possible to carry out the visits in spring 2022, there would have been interested applicants for many more visits. There clearly was a need for this kind of activities and it is a shame that the pandemic limited the amount of visits to such a degree.

The fact that we received several applications from countries not associated with EU, which we could not offer a grant, shows an interest for the ELEXIS transnational activities that goes far beyond EU and Europe.

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, the network building and knowledge exchange are very important elements of the success of the visits. 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.

The fact that several grant holders now are writing papers, giving talks at conferences or have started on a PhD project is another proof of the success of the transnational activities.

Many of the hosts report that they too have been very happy to host visits, that they have benefited from the knowledge exchange and have extended their network correspondingly by getting to know lexicographers and researchers from other institutions working in parallel fields.

Appendix A

Below are attached the reports from the last year. The reports are from the final call as well as from postponed visits from earlier calls and include the following visitors:

- Denis Gaščić: Methods for detection and evaluation of neologisms for the Croatian language.
- Eglantina Gishti: A Corpus-based method for Extraction of Polylexical Units (in French and Albanian languages).
- Ana Mihaljević: Dictionary of the Croatian Redaction of Church Slavonic: exploring the possibilities of digitalization, retrodigitalization & modernization of the dictionary writing process.
- Dorota Mika: Integration of lexicographic data: the diachronic plane
- Carolin Müller-Spitzer: How is the corpus influencing collocation sets in dictionaries? Enhancing a study on German collocation sets for Mann and Frau to a contrastive German-Dutch study.
- Annika Simonsen: Ravnur – the Faroese Speech Recognizer
- Dóra Mária Tamás: The Description of the Characteristics of Legal Terminology and Issues of Editing Legal Databases in Lexicography and Terminology.

One of the reports are still in progress and is therefore not included here but will be uploaded to our website once it is completed: Jelena Parizoska: Adapting dictionary writing systems and other platforms to online dictionaries of idioms.

Denis Gaščić: ELEXIS visit grant

Methods for detection and evaluation of neologisms for the Croatian language

QUESTIONNAIRE BEFORE THE VISIT

How did you learn about the ELEXIS travel grants?

I was a MA student of information science at the Faculty of Humanities and Social Sciences at the University of Zagreb when Call 4 was announced. My professor Kristina Kocijan sent me a call, so I decided to apply because I am interested in natural language processing and e-lexicography.

What is your project about?

The purpose of my visit is to obtain an overview of

1. methods and tools for detecting and evaluating neologisms;
2. Estonian resources (corpora, lexical resources) and tools used for (semi-)automatic detection of neologisms;
3. Croatian resources (corpora, lexical resources) and tools that can be used or are used for (semi-)automatic detection and evaluation of neologisms.

During the visit, I also plan to define a theme and write a research proposal for my PhD thesis, probably in the field of neology.

What is your background that brought you up to this point?

I have a MA in information science, but I would say that I am a computer and language enthusiast who every day tries to learn some new things and acquire new skills.

Which hosting institution did you apply to and why?

The hosting institution, Institute for the Estonian Language, is focused on modern automated lexicography, so they can introduce me methods and tools (DWS-s and CQS-s) they use for dictionary compilation and automatic detection of neologisms. They can also teach me how to use them and implement them in the research of the Croatian language.

Where does your interest in lexicography come from and what keeps you motivated?

I was always interested in creating new words (I created a few of them, and two of them were chosen for the final competition for new Croatian words). I am also interested in the whole lifecycle of neologisms. My area of interest also includes the field of machine translation, speech recognition, spell checking, diachronic analysis and automated detection of neologisms. I hope I will be able to develop algorithms for the automatic detection of neologisms for Croatian. My primary motivation is improving and developing NLP applications for modern Croatian and their practical implementation in tools and apps.

Travel Grant	Call 4	
Period of stay	12. – 26.6.2022	
Project title	Methods for detection and evaluation of neologism for the Croatian language	
Home institution	Faculty of Humanities and Social Sciences, Zagreb (Filozofski fakultet, Zagreb)	#elexis_hr
Hosting institution	Institut of the Estonian language (Eesti Keele Instituut)	#elexis_ee

****

REPORT

The period of my visit was from the 12th to the 26th of June. During the first week, I met colleagues from the Institute of the Estonian Language and participated in the 19th Conference of Applied Linguistics, “Influence of the language: from Data to Content-Rich Knowledge”, organised by the Estonian Association of Applied Linguistics. During the second week, I focused on my research and studied bibliographic sources. finalised my project result and wrote a PhD proposal.

1. Introduction

During a project, I obtained an overview of

1. methods and tools for detection and evaluation used in modern lexicography (e.g. Sketch Engine, Neoveille, Google Ngram Viewer);
2. Estonian resources (corpora, taggers, and lexical resources) used for (semi-) automatic detection of neologisms
    - lexical databases: ekilex.ee, WordNet;
    - corpora: Estonian National Corpus 2021, incl.Web Corpus 2021 and monitor corpora Timestamped_Feeds_2014-2021;
    - tools for Estonian NLP: NLP Toolkit for Estonian Estnltk and Universal Dependencies for the Estonian language;
    - Corpus Query Systems: Sketch Engine, Korp
3. Croatian resources (corpora, taggers, tools and lexical resources) that can be used or are used for (semi-)automatic detection and evaluation of neologisms
    - corpora: National Corpus, corpora in Sketch Engine;
    - lexical databases and dictionary portals: Hrvatski jezički portal, Hrvatski jezični korpus, Mrežnik;
    - tools for Croatian NLP: Universal Dependencies for Croatian.

I also defined a topic and wrote a research proposal for my PhD thesis. The subject of my PhD thesis will be “Grammar checker for the Croatian language: theory and modelling”.

For implementing a grammar checker, there are many preconditions to be done. One of them is also automatic detection of neologism because it is essential to distinguish between real neologisms and misspelt words. This is also why I, during the research visit, focused on the automatic detection of neologisms. Furthermore, besides detecting neologism, for developing a grammar checker, it is necessary to have corpora, structured lexical data (which includes the misspelt words related to right-spelt) and the API, which connects the database with the application. However, much time is needed to explore all these preconditions, so I primarily focused on neologism.

2. Description of work carried out during the research visit

On the first day of the project (June 13), I met my host Jelena Kallas, a Senior Computational Lexicographer-Project Manager at the Institute of the Estonian Language. She familiarised me with the Institute and its work.

During our meeting, we discussed primarily tools (DWSs and CQSs) used for dictionary compilation and neology detection (see part 2). The Institute of the Estonian Language uses DWS ekilex.ee and CQSs Sketch Engine and Korp. There is also a special NLP Toolkit for Estonian Estnltk, a Python library for performing common language processing tasks in Estonian. This toolkit is developed at the University of Tartu.

The next day (June 14), we met Martin Luts, a machine translation expert from the Institute. He gave me insight into new machine translation technologies, especially in using neural network algorithms and combining them with other technologies like statistical machine translation (SMT). We also discussed the importance of human feedback and correct training methods. Finally, we noticed that the question arises: "Where to store the data and does, for security reasons, text with confidential information can be translated via commercial translation apps?". I also met the Institute's NLP engineer Silver Vapper, who consulted me about optical character recognition (OCR) and bilingual lexicography. Although Estonia is a highly digitised country, OCR is needed to digitise old texts, i. e. to add their content to corpora, which is vital to see trends with words. In Croatia, on the other side, OCR

is also needed for getting modern language corpora too, because the government and public institutions still produce paper-based content.

On the third day (June 15), I met the Institute's project manager Marja Vaba. We discussed the Institute's products, especially Ekilex and Sõnaveeb. Sõnaveeb is the language portal of the Institute containing linguistic information from a growing number of dictionaries and databases. We concluded that one of the most important things is to know what the user of the language portal wants and what he/she needs, especially if he/she is not able to specify his/her needs. For that, we concluded that it is essential to research users' habits and needs, but also e.g. their educational background.

Maybe the most helpful conversation for my project was a discussion with Istok Kosem, a research assistant from the University of Ljubljana, who also visited the institute. Iztok is an e-lexicographer and NLP expert for the Slovene language. Because of the similarities between Croatian and Slovene, I could draw parallels between language technologies used for Slovene and those needed for Croatian. Istok also presented me for Sloleks (Slovenski oblikoslovni leksikon), Slovene Morphological Lexicon, based on the thesaurus database available at the CLARIN.SI repository. Sloleks is a lexicon of Slovene word forms, containing 100,802 headwords and 2,792,003 word forms with grammatical and accentual features. The innovation that Slolex offers to the user, compared to other lexicons, is that it predicts what the user wants when he is still typing. More precisely, it means that the application customises his/her search menu, so it offers not only a word but also additional information (e. g. the type of word, gender, grammatical information).

On Thursday and Friday (June 15-16), the fourth and fifth days of my visit, I participated in the 19th Annual Conference of Applied Linguistics. The programme included presentations about the most modern technologies and tools for language processing, but presentations about current unresolved issues in (corpus) lexicography.

During the second week (June 20-25) of the visit, I was focused on studying bibliographic sources (I used Elexifinder, EURALEX Proceedings, eLex Proceedings, materials of Globalex workshops on Lexicography and Neology), analysis of the corpora (mostly web corpora and monitor corpora) available for Croatian, finalizing my project result and writing a PhD proposal.

3. State-of-the-art techniques in neography

3.1. Introduction

According to McEnery, Xiao & Tono (2006), a corpus is "a collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety". Nevertheless, one of the main characteristics of language is that it is changeable. The changings of the language from the corpora perspective, in the opinion of the author of these lines, can be observed in three ways: (1) as changing by topics, (2) as changing in space or group, and (3) as a changing by time.

Firstly, when two people are talking about sports, they do not use the same vocabulary as two people talking about politics. Even the way they construct sentences can be different. In addition, the language on the same territory is different from the language on the other territory. At the same time, the language of the young is different from the language of the older people. There also can be some subcultures that use their part of language, which we call slang. In linguistics, those parts of language are known as dialects. According to the Cambridge Dictionary, “dialect is a form of a language that people speak in a particular part of a country, containing some different words and grammar”.

For those reasons, the defining feature of a corpus is representativeness. Leech (2011) defines a corpus as representative "if the findings based on its contents can be generalized to the said language variety".

Finally, a language is changeable over time. Some words are coming, while the others, at the same time, disappear. The words which come up are called neologisms. As for anything in humanities and social sciences, there is no one definition for neologisms. Plag explains them as “those derivatives that were newly coined in a given period” (Plag, 2002).

3.2. Neologisms and their detection

According to Cartier (2017), for tracking neologism in corpora, what is needed is "to have at hand large diachronic electronic corpora". Nevertheless, there are more problems with neology. Firstly, linguistics still observes neology as a non-primary field, as Cartier (2017) pointed out. On the other hand, Klosa and Lüngen (2018) claim that “a central issue in

lexicography is to find new lexemes and to identify new meanings for existing lexemes". Nevertheless, it is not easy to observe neologism automatically because computers do not understand meanings. So, the first way to trace neologisms was to induce unknown words. That system is called "exclusion dictionary architecture" (EDA) and includes two parts: monitor corpora and a list of known words with a list of misspelt words (Ibid.). The problem is that neologism is not only a new word, sometimes the word, which already exists, can get a new meaning. For example, the word "mouse", with the advent of computers, got meaning "the part of the computer for navigation on-screen". Therefore, another type of existing neology tracking system is Semantic Neology Approaches. It is based on the principle of computational understanding of contexts. Behind it is the "idea that meaning change is linked to domain change: every text and thus the constituent existing lexical units are assigned one or more topic; if a lexical unit emerges in a new domain, a change in meaning should have occurred (Gerard et al., 2014)". Based on that, Cartier created a Neovelle web platform for neologism tracking, which architecture has five components: a corpora manager, an advanced search engine on the corpora, advanced data analytics, a linguistic description component for neologisms and formal and semantic neologisms tracking with state-of-the-art techniques (Cartier, 2017).

In a Sketch Engine, there is a Trends function, which gives users (e.g. lexicographers) the opportunity to observe not only the appearance of new words but also the disappearance of old ones. It is based on mathematical methods, which are calculated by Python script developed by Ondřej Herman (2013). Heman was also comparing a lot of mathematical methods for corpus changing monitoring. He concluded that: "the method with the highest cost-to-benefit ratio for implementation seems to be the Theil-Sen slope estimator, along with the Spearman's ρ or Mann-Kendall tests to investigate a possible trend present in the word usage data" (Ibid.). It is, he described, "calculated by the attached code along with the implementation of the other regression methods. The code interfaces with Sketch Engine and can read data from Google n-grams datasets" (Ibid.).

Because Google has a large amount of text obtained by digitising books and tracking web content in English, it has enabled its users to track changes in language. The Google n-gram dataset is "a publicly available corpus with co-occurrence statistics of a large volume of web text" (Koplenig, 2017). An N-gram is, as Mazumder, Sourav and Baru (2022) pointed out, "a contiguous sequence of n words or tokens in a text document in computational linguistics and probability". It is a probabilistic language model that "can be classified into categories depending on the unit that incorporated them" (Ibid.).

In 2018, a group of authors in the context of the Horizon 2020 project ELEXIS surveyed lexicographic practices and lexicographers' needs across Europe. The results have shown that only 4,7 % of lexicographers across Europe use automatic extraction of neologisms. The research also shows that "the majority of the respondents compile their dictionaries manually (57.9%)" (Kallas et al., 2019.).

3.3. Resources for Croatian: corpora and lexicon

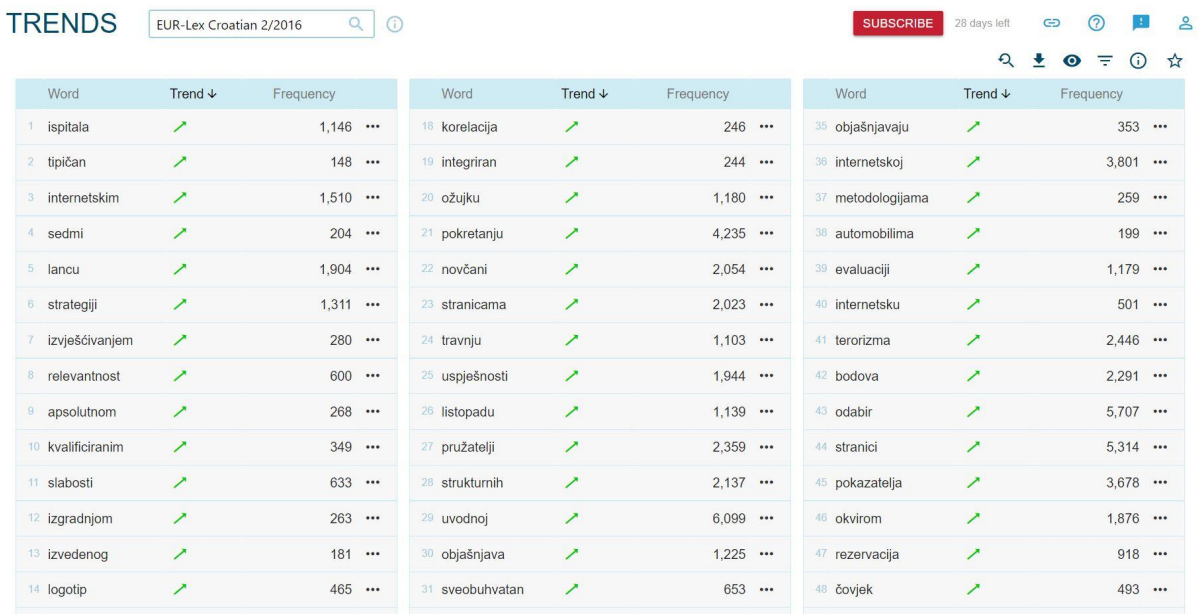
For the Croatian language, unfortunately, there are not so many corpora available. The Croatian National Corpus (Hrvatski jezični korpus), the largest one, is collected at the initiative of Prof Marko Tadić from The Faculty of Humanities and Social Sciences, and has 2,559,160 words and 2,130,095 lemmas. The last time it was updated was 11/02/2021, which is one and a half years before writing these lines. It is also based on NoSketchEngine, a free version of Sketch Engine that does not provide all the features. In addition, there is also Croatian Web (hrWaC 2.2, RFTagger) corpus in SketchEngine, which has 1,405,794,913 tokens and 1,211,328,660 words. It was crawled in 2011 and 2013, so it does not provide users with the real state of the language.

From the dictionary perspective, only the Hrvatski jezički portal (The Croatian Language Porta, HJP) is available. The Croatian Language Portal is the first and so far the only dictionary database of the Croatian language distributed on the Internet, which has been available free of charge since June 2006. The project received initial support from the Ministry of Science, Education and Sports in 2004 and has since been funded by the owner's funds. The Croatian Language Portal is the only such scientific reference work in Croatia. This dictionary requires continuous, detailed and painstaking work of several experts in the field of linguistics and other social sciences and humanities scientists to be updated following current knowledge and constant enrichment base. Unfortunately, for example, it does not contain coronavirus terms, which means that it has also not been updated for more than two years, as these words remain in the language.

Based on the Croatian Web Repository, there is also a Croatian Web-Dictionary – Mrežnik project. Authors say that "Croatia still belongs to the ever-smaller number of countries with no free online national language dictionary founded on modern e-lexicography, nor has systematic scientific research been carried out in this area" (Mrežnik). So, "the basic goal of this project is to change this in both of the aforementioned aspects.". The project is still in the working phase.

At the 3rd Globalex Workshop on Lexicography and Neology in 2021, Mihaljević, Hudeček and Lewisis (2021) presented a paper "Corona-related neologisms: A challenge for Croatian standardology and lexicography". Their research was also based on manually collected corona-terms because there was no automatic, even semi-automatic, system for tracing neologisms.

Sketch Engine has a few more corpora for the Croatian language. Only one enables the Trends function. That is a EUR-Lex Croatian 2/2016, EUR-Lex multilingual corpus of all the official languages of the European Union, which contains (only) 17,819,540 sentences and 156,309,317 words. Unfortunately, the quality of lemmatization and morphological analysis is not good enough (more detailed evaluation is needed), there are a lot of mistakes. It might be good to evaluate the lemmatizer used by Sketch Engine and possibly use another one developed especially for Croatian, not for Slovene. As a result, when we search trends by lemmas, we get almost the same result as searching by words (compare figures 1 and 2).

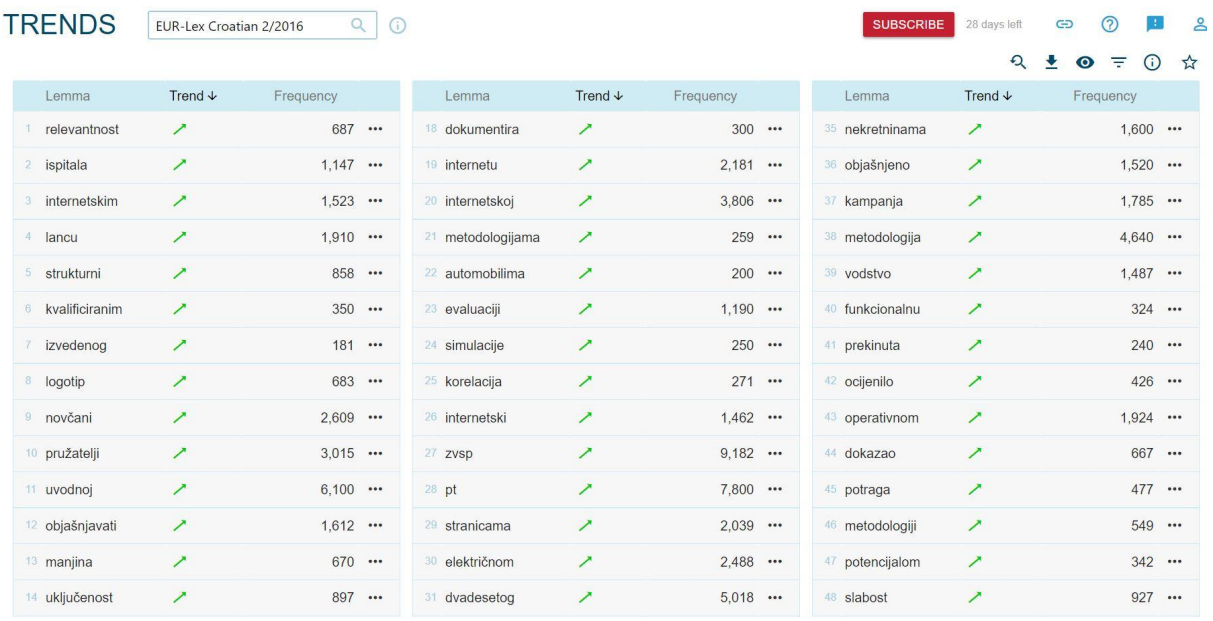


	Word	Trend	Frequency
1	ispitala	↗	1,146
2	tipičan	↗	148
3	internetskim	↗	1,510
4	sedmi	↗	204
5	lancu	↗	1,904
6	strategiji	↗	1,311
7	izvješćivanjem	↗	280
8	relevantnost	↗	600
9	apsolutnom	↗	268
10	kvalificiranim	↗	349
11	slabosti	↗	633
12	izgradnjom	↗	263
13	izvedenog	↗	181
14	logotip	↗	465

	Word	Trend	Frequency
18	korelacija	↗	246
19	integriran	↗	244
20	ožujku	↗	1,180
21	pokretanju	↗	4,235
22	novčani	↗	2,054
23	stranicama	↗	2,023
24	travnju	↗	1,103
25	uspješnosti	↗	1,944
26	listopadu	↗	1,139
27	pružatelji	↗	2,359
28	strukturnih	↗	2,137
29	uvodnoj	↗	6,099
30	objašnjava	↗	1,225
31	sveobuhvatan	↗	653

	Word	Trend	Frequency
35	objašnjavaju	↗	353
36	internetskoj	↗	3,801
37	metodologijama	↗	259
38	automobilima	↗	199
39	evaluaciji	↗	1,179
40	internetsku	↗	501
41	terorizma	↗	2,446
42	bodova	↗	2,291
43	odabir	↗	5,707
44	stranici	↗	5,314
45	pokazatelja	↗	3,678
46	okvirom	↗	1,876
47	rezervacija	↗	918
48	čovjek	↗	493

Figure 1. The Trends search by words



	Lemma	Trend ↓	Frequency
1	relevantnost	↗	687
2	ispitala	↗	1,147
3	internetskim	↗	1,523
4	lancu	↗	1,910
5	strukturni	↗	858
6	kvalificiranim	↗	350
7	izvedenog	↗	181
8	logotip	↗	683
9	novčani	↗	2,609
10	pružatelji	↗	3,015
11	uvodnoj	↗	6,100
12	objašnjavati	↗	1,612
13	manjina	↗	670
14	uključenost	↗	897

	Lemma	Trend ↓	Frequency
18	dokumentira	↗	300
19	internetu	↗	2,181
20	internetskoj	↗	3,806
21	metodologijama	↗	259
22	automobilima	↗	200
23	evaluaciji	↗	1,190
24	simulacije	↗	250
25	korelacija	↗	271
26	internetski	↗	1,462
27	zvsp	↗	9,182
28	pt	↗	7,800
29	stranicama	↗	2,039
30	električnom	↗	2,488
31	dvadesetog	↗	5,018

	Lemma	Trend ↓	Frequency
35	nekretninama	↗	1,600
36	objašnjeno	↗	1,520
37	kampanja	↗	1,785
38	metodologija	↗	4,640
39	vodstvo	↗	1,487
40	funkcionalnu	↗	324
41	prekinuta	↗	240
42	ocijenilo	↗	426
43	operativnom	↗	1,924
44	dokazao	↗	667
45	potraga	↗	477
46	metodologiji	↗	549
47	potencijalom	↗	342
48	slabost	↗	927

Figure 2. Trends search by lemmas

The word “ispitala” (Croat. “examined”) has been shown as the most frequent lemma. It is an impersonal participle form of the verb “ispitati”, but it is lemmatised as a masculine noun. In addition, there are also tokenisation mistakes. The problem is that for tokenisation Sketch Engine uses the MULTEXT-East Slovenian part-of-speech tagset. Although Croatian and Slovenian are similar languages, it is still reasonable to use taggers developed especially for Croatian, even quite frequent adjectives and numerals have wrong lemmas and POS. On the other hand, the word "ispitala" is in the Croatian Web (hrWaC 2.2, ReLDI) corpus tokenised well - as the verb participle singular feminine. For the lemmatization of the Croatian Web (hrWaC 2.2, ReLDI) the MULTEXT-East Croatian part-of-speech tagset is used. That tagset is a product of the MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages), whose main task it was to develop standardised language resources (Erjavec et al., 2017). The Croatian specifications were compiled soon after the MULTEXT-East project ended in 1997, using the project's Final report as the template (Ibid.). One of the objectives of MULTEXT-East has been to make its resources freely available for research (Ibid.). So, after my visit, I plan to contact the Sketch Engine team and propose to use a different parser for Croatian.

4. Conclusion remarks

Detection of neologism is a field which needs further research. A lot of lexicographers nowadays still detect neologisms manually, which takes time that they could devote to other tasks. As Kilgarriff et al. (2015) describe, “lexicographers read texts which are likely to contain neologisms – newspapers, magazines, recent novels – and mark up candidate new words, or new terms, or new meanings of existing words. It is a high-precision, low recall approach, since the readers will rarely be wrong in their judgments, but cannot read everything, so there are many neologisms that will be missed” (Ibid.). Only automated methods for corpus linguistics can provide a systematic analysis of large amounts of text, offering neologism candidates to lexicographers. There is a need to set up an infrastructure for neologism detection to supply lexicographers working with neologisms with candidates for inclusion in dictionaries. The problem with the detection of neologism is how to recognize not only new words but also new meanings of words which already exist. That is a reason why it is vital to also develop semantically annotated corpora, develop algorithms for sense clustering and share expertise in this field. This issue has been dealt with in the ELEXIS project, see e.g. deliverable on the topics of semantically annotated corpora, and Word Sense Disambiguation (WSD) algorithm for sense clustering, developed by Federico Martelli and Roberto Navigli (the results are available also at GitHub). They conclude that there are also two more directions of text analysis to have been explored: domain-labelling of texts and diachronic distribution of senses (Martelli et al., 2019).

On the other hand, there is no predisposition to implement (semi-)automatic detection of neologisms in the Croatian language nowadays. However, to enable such a system, a few essential things must have been done. Firstly, it is necessary to have a big, timestamped corpora so that language changes can be followed regularly, for example, monthly. It is also essential to develop monitor corpora and create Web corpora. A good example for Croatia can be Slovenia. For example, the newest version of Slovenian corpus Trendi (version 2022-05) contains 565.308.991 lemmas from 1.436.548 words. The Trendi 2022-05 corpus is available in three CLARIN.SI concordances: KonText, NoSketchEngine and the old version of the NoSketchEngine interface.

Mentioning the web corpora, there are two main problems with them. The first is cleaning, which means "removing those sections of a document that are textual but not linguistically informative" (Pomikálek, 2011), such as advertisements, headers, etc. The second problem is removing duplicate text (Ibid.) so the system is representative.

Furthermore, developing an advanced search engine and NLP tools for the corpora is essential. In addition, it is necessary to give human expert feedback to the system. Ultimately, it is crucial to follow the state-of-the-art technology and regularly analyse which methods and tools are used for other languages, especially Slavic, which could be implemented for the Croatian language. Without these predispositions (lemmatized corpora, dictionaries, thesaurus databases), there is also no predisposition for developing a (functional) grammar checker.

In summary, it is essential, as Tiberius et al. (2020) pointed out, to create "robust documentation, guidelines and collections, best practices in order to promote clearly defined workflows for producing, describing and annotating lexicographic resources (both synchronic and diachronic) in accordance with international standards and interoperability formats" (Ibid.).

The main problem with research of the Croatian language is, in the opinion of the author of these lines, that language processing is not recognized as an essential field, resulting in the non-investment of public money in language technologies. This field is also not recognized in the private sector. In my opinion, it is time to change the language policy in Croatia and to start investing in the development of language technologies.

Lastly, I would like to repeat that text analysis is one of the fields that still have to be discovered, especially when we talk about detecting neologisms. Although the lack of tools for automatic detection of neologisms is a problem today, at the same time, it could also be an opportunity for researchers like me who are interested in developing new things. Because of all that has already been said, I can find myself doing a PhD thesis in natural language processes.

5. Final words

The research visit grant at the Institute for the Estonian language played a significant role in my understanding of neology, automation detection of new words, but also in natural language processing generally. It was also useful and enjoyable to participate in the 19th Annual Conference of Applied Linguistics to get a professional overview of state-of-the-art methods and tools.

Furthermore, the themes and authors I have discovered during my visit to the Institute and the conference inspired me to continue with new research in the field. At the

same time, it inspired me to implement similar tools for the Croatian language because I consider how intensive and efficient their role is, especially in practical use by native speakers but also by language learners. .

Resources developed and used at Institute are practical, and they support both lexicographers and other language-orientated scientists like sociolinguists to get a better understanding of language, language processes and their social impacts. In addition, I would like to stress that NLP tools developed by the University of Tartu and the Technology University of Tallinn are open-sourced, which means they could be reused and implemented for other languages, like Croatian.

Finally, I was honoured to have been invited to the Institute of the Estonian Language to meet NLP engineers, lexicographers and other people who work at the Institute, especially my host Jelena Kallas. Of course, the knowledge I got from them will help me in my future work, but I think it is more important that their work inspired me.

References

Publications

1. Arppe, A. (2000, December). Developing a grammar checker for Swedish. In *Proceedings of the 12th Nordic Conference of Computational Linguistics* (NODALIDA 1999) (pp. 13-27).

2. Cartier, E. (2017, April). Neoveille, a web platform for neologism tracking. In Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (pp. 95-98).

3. Erjavec, T., Krstev, C., Petkevic, V., Simov, K., Tadić, M., & Vitas, D. (2003, April). The MULTEXT-East Morphosyntactic Specification for Slavic Languages. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages* (pp. 25-32).

4. Herman, O. (2013). Automatic methods for detection of word usage in time: Bachelor's thesis. Masaryk: Masaryk University, Faculty of Informatics.

5. Kallas, J., Koeva, S., Langemets, M., Tiberius, C., & Kosem, I. (2019, October). Lexicographic practices in Europe: Results of the ELEXIS Survey on user needs. In Electronic Lexicography in the 21st Century. In *Proceedings of the eLex 2019 Conference*, Sintra, Portugal (pp. 1-3).

6. Kilgarriff, A., Herman, O., Bušta, J., Kovář, V., & Jakubíček, M. (2015, August). DIACRAN: a framework for diachronic analysis. In *Proceedings of Corpus Linguistics* (pp. 65-70).

7. Klosa, A., & Lüngen, H. (2018, August). New German words: Detection and description. In *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts 17-21 July 2018*, Ljubljana (pp. 559-569). Znanstvena založba Filozofske fakultete Univerze v Ljubljani/Ljubljana University Press, Faculty of Arts.

8. Koplenig, A. (2017). The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. Digital Scholarship in the Humanities, 32(1), 169-188.

9. Leech, G. (2011). Corpora and theories of linguistic performance. In *J. Svartvik (Ed.), Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm,* 4-8 August 1991 (pp. 105-126). Berlin, New York: De Gruyter Mouton.
10. Liou, H. C. (1991). Development of an English grammar checker a progress report. CALICO Journal, 57-70.
11. Martelli, F.; Navigli, R; Spadoni, P.; Stilo, G.; Velardi, P. (2019). Lexical-semantic analytics for NLP: sense clustering. ELEXIS - European Lexicographic Infrastructure.
12. McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. Taylor & Francis.
13. Mihaljević, Hudeček, Lewisis. (2021). Corona-related neologisms: A challenge for Croatian standardology and lexicography. At: Globalex Workshop on Lexicography and Neology, 2021. Virtual. (Presentation).
14. Mikkelsen, I. L. S., Wiechetek, L., & Pirinen, F. A. (2022, May). Reusing a Multi-lingual Setup to Bootstrap a Grammar Checker for a Very Low Resource Language without Data. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 149-158).
15. Plag, I. (2002). Word-formation in English. Cambridge: Cambridge University Press.
16. Pomikálek, J. (2011). Removing boilerplate and duplicate content from web corpora. Disertacnı práce, Masarykova univerzita, Fakulta informatiky.
17. Tiberius, C., Costa, R., Erjavec, T., Krek, S., McCrae, J., Roche, C., & Tasovac, T. (2020). Best practices for lexicography – intermediate report. ELEXIS - European Lexicographic Infrastructure.

Other resources

1. CLARIN.SI
2. Ekilex
3. Estnlt
4. Estnltk
5. Google Ngram Viewer
6. Hrvatski jezički portal

7. Hrvatski jezični korpus
8. Korp
9. Mrežnik
10. Neoveille
11. Sketch Engine
12. Sloleks
13. Sõnaveeb
14. Universal Dependencies for Croatian language
15. Universal Dependencies for the Estonian language
16. WordNet

Eglantina Gishti, Professor of Linguistics
University of Tirana
Tirana, Albania

Report on Elexis Transnational Research Visit Grant at Det Danske Sprog- og Litteraturselskab and at the Centre for Language Technology, Department of Nordic Studies and Linguistics, University of Copenhagen (Copenhagen, Denmark, April 4 – April 8, 2022)

Project title:

A Corpus-based method for Extraction of Polylexical Units

(in French and Albanian languages)

Introduction

I applied for a visit to Det Danske Sprog- og Litteraturselskab and to the University of Copenhagen to discuss various aspects of the work related to dictionaries and corpora and to know the functions of all the corpus they created and their specificities. Since my research concerns lexicography and NLP and being aware of the lack of lexicography work and products in Albania, I wanted to visit an institution where a lexicographic project is being conducted (new lexicographical methodology) in order to exchange ideas and discuss different issues concerning the web dictionary. In that point of view, at Det Danske Sprog- og Litteraturselskab (DSL) I was introduced with the project ordnet.dk, a comprehensive corpus-based dictionary of modern Danish and other important projects ongoing in Det Danske Sprog- og Litteraturselskab (DSL).

Furthermore, I visited the Centre for Language Technology at the University of Copenhagen, met the researchers, and was introduced to their projects, tools and resources. My initial idea was to conduct research, in which we will create parallel data on foreign languages and Albanian language that will allow us to do research and to establish a rich database for the language technology. The visit at the University of Copenhagen made me aware that it will be more convenient to start with a monolingual corpus because this will enable us to have more complete data in the respective languages. Once it is done, the work on parallel corpora will be easier. Based on the experience in both institutions, I realized the importance of the lexicon represented in corpora and dealt with in dictionaries. In my project interest, this is important as we work as well

with the Extraction of Polylexical Units that are difficult to be presented in a dictionary. The reference to the Corpora is a must in that case.
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 of the Centre for Language Technology at the University of Copenhagen. Besides that, everyone was very helpful and eager to answer my questions.

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

4 – 5 April 2022: visit to the Centre for Language Technology (University of Copenhagen)

During my stay in Copenhagen, I started the visit at the Centre for Language Technology (Center for Sprogteknologi) at the University of Copenhagen where I was welcomed by Prof. Bolette Sandford Pedersen and Sussi Olsen. They presented me to the other staff members and the researchers who introduced me to their main projects, among which are the following:

- *CLARIN-DK* as an infrastructure where researchers can deposit, share, and download language-based material. i.e., texts, transcriptions, lexicons, word lists, audio, and video files. CLARIN-DK also comprises interactive language tools.
- *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.
- *The Danish FrameNet* – based on the Berkeley FrameNet model
- *ParlaMint* - a project which contributes to the creation of comparable and uniformly annotated multilingual corpora of parliamentary sessions. This was interesting w.r.t. the idea of parallel corpora idea.
- ONP: Dictionary of Old Norse Prose - http://onp.ku.dk/
- Wordnet tool
- Web scraping methods

Furthermore, the researcher at the Centre for Language Technology offered to start **training CLARIN tools f**or Albanian.

6-8 April 2022 - Scheduled meetings at Society for Danish Language and Literature

During my visit, I spent the rest of my mobility at the Society for Danish Language and Literature, and I was welcomed there by Sanni Nimb. 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) and *Den Danske Begrebsordbog* (a Danish thesaurus)

- I was also introduced also to Ordbog over det danske Sprog (ODS; a dictionary of older Danish published online at ordnet.dk/ods), because of my interest in historical texts – not only lexicography and modern corpora.
- Concerning the building of monolingual corpora: I discussed with Ida Flörke the issues of collecting text material from the publishing houses and she showed how the juridical contract is formulated. She also provided me with a version in English.
- introduction to the **tools** used by the lexicographers at DSL:
    - dictionary writing system iLex
- 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 (**tool for linguistic studies** in very large text collections).
        - frequency word lists generated from the corpora
        - 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 define.
- **other resources** developed at DSL:

I also met the director of DSL, Dr. Karen Skovgaard-Petersen and Finn Gredal Jensen. They showed me some of their work. One of these works is the online Holberg edition: http://holbergsskrifter.dk. And we exchanged experiences on our current work.

Dr. Marita Akhøj Nielsen told me about Leonora Christina's French autobiography and Music and language in the Danish hymn singing of the Reformation period.

In addition

The last day of my mobility, we had a visit at the Royal Library, guided by Dr. Anders Toftgaard. This visit was possible because DSL collaborates closely with the library. We profited to see among different collections: the Royal Library's Manuscript Collection contains manuscripts which range from the early Middle Ages to the present. We were introduced to a lot of information concerning the library, its resources, and its history.

We also discussed the lexicographical situation of the Albanian language, my work on corpora and dictionaries. We exchanged different ideas about how to improve the lexicographical work, about the importance of the books and manuscript's digitization, etc.

Conclusion

The visit proved to be above my expectations. From historical data and dictionaries to different tools, the projects I was introduced revealed to be a true inspiration and to rethink the methodology I had conceived for my project. I believe that the experience and contacts I have gained from it will be very valuable for my work. 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 the tools and Danish lexical resources. Through informative and inspirational conversations with the editors of the Danish dictionary and the researchers, I was provided with ideas for my own work.

I would like to use this occasion to thank my hosts – Centre for Language Technology (University of Copenhagen) and 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 researchers and editors at both institutions for explaining to me all their work and answering all my questions. I would, in particular, like to express my thanks to Dr. Sanni Nimb, Prof. Bolette Sandford Pedersen and to Sussi Olsen, for assisting me prior to and during my visit as well as planning my activities and for administrative assistance. Finally, thank you to the Elexis project for making this research mobility possible.

11/4/2022

dr. sc. Ana Mihaljević, research associate
amihaljevic@stin.hr
Old Church Slavonic Institute
Zagreb
Croatia

ELEXIS visiting grant report – Institute of Bulgarian Language Lyubomir Andreychin

From May 7 to 21, I was on an ELEXIS visiting grant at the Institute of Bulgarian Language. The main goal of my visit was to explore the possibilities of digitalization, retrodigitalization and modernization of the dictionary writing process since I am one of the lexicographers working on the compilation of the dictionary of the Croatian Redaction of Church Slavonic. This dictionary is the first dictionary of Croatian Church Slavonic, and it is currently being compiled using MS Word and is, for now, available only in the physical form. The corpus for this dictionary consists of paper handwritten card files that are currently in the process of OCR. During my visit to the Institute of Bulgarian Language Lyubomir Andreychin I met with various lexicographers and computer linguists, mainly working on the Dictionary of the Bulgarian language. It is the most extensive, most comprehensive, and most representative monolingual dictionary of the Bulgarian language. So far, 15 volumes have been published with entries from a to r and include more than 100 000 entries.

On Monday, May 9, I met with the head of the Department of Bulgarian Lexicology and Lexicography prof. Diana Blagoeva, the head of the Department of Computational Linguistics prof. Svetla Koeva and a lexicographer from the Department of Bulgarian Lexicology and Lexicography Sia Kolkovska. They presented their work on the Dictionary of Bulgarian Language and the work of their departments and gave me a fundamental insight into the compilation process. They also showed me their physical corpus and explained its digitization process.

On Tuesday, May 10, I met with assist. prof. Tsvetana Ivanova Dimitrova and Valentina Georgieva Stefanova, who introduced me to the Bulgarian sources for historical languages (historical dictionaries, corpora, programs for syntactic linking of the translational texts, etc.) and the work on the Etymological dictionary.

On Wednesday, May 11, I met the entire Department of Bulgarian Lexicology and Lexicography, presented my work and some of the problems with retrodigitization of the corpus and previously published volumes and the compilation process, and learned more about their work.

On Thursday, May 12, I met with Michaela Kuzmova, a lexicographer from the Department of Bulgarian Lexicology and Lexicography, who showed me the possibilities of the online version of their dictionary and with whom I discussed some of the problems concerning the digital version of the dictionary. In the evening, I participated at the conference celebrating the hundredth anniversary of the Department of Classical Philology at the University of Sofia.

On Friday, May 13, I met with computer scientist Borislav Stankov who showed me how to work in the dictionary writing system LexIt, which he developed for the Institute. We discussed the advantages and disadvantages of this program and other dictionary writing systems. I also tried to use this program for my dictionary project.

From Sunday, May 15 to Tuesday, May 17, I attended the annual conference of the Institute. I met with many researchers working in different fields and exchanged experiences at the conference. In the evenings, I met with assist. prof. Simeon Stefanov from the Department of Bulgarian Etymology, with whom I discussed some problems in their dictionary compilation process and some problems and challenges in the field of historical linguistics. I also met with prof. Margaret Dimitrova, Iskra Hristova Shomova, and Aneta Dimitrova from the University of Sofia, Ekaterina Dikova and Lora Taseva from the Institute of Balkan Studies and Centre of Tracology, and Petra Stankovska from the University of Ljubljana and discussed various lexicographic problems relevant for our project.

On Wednesday, May 18, I met with Michaela Kuzmova, who showed me different types of corpora and how they use various programs for corpus search.

On Thursday, May 19, I held an online lecture entitled *Croatian Church Slavonic in the Digital Era* at the Spring linguistic seminar.

On Friday, May 20, I worked at the library and tried using some of the ELEXIS programs, such as Elexifier, to convert my dictionary, Lexonomy, for future dictionary compilation and publishing. I also learned about NAISC and EDiE, which I learned even further about in the online lecture *ELEXIS: AI-powered tools for lexicographers in the 21st century* on May 25.

During my visit to the Institute of Bulgarian Language, I learned a lot about different dictionary writing systems (such as LexIT, tLex, iLex, Lexonomy). I discussed their usage with more experienced lexicographers and computer scientists and tried using some of the programs for my dictionary project. I also learned about different programs used to facilitate managing different types of corpora. I think this visit will enhance my work. During the visit, I learned

to use various programs and learned about different lexicographic projects and gained experience that will shifted the way I plan my future work in this field.

ELEXIS TRANSNATIONAL RESEARCH VISIT GRANT FINAL REPORT

Grant holder: Dorota Mika, PhD

Affiliation: Institute of Polish Language, Polish Academy of Sciences (Poland)

Hosting institution: Instituut voor de Nederlandse Taal (Netherlands)

Period: 30.05.2022-3.06.2022

Project Title: Integration of lexicographic data: the diachronic plane

Research objectives

The main problem currently confronting linguists using lexical resources is the significant fragmentation of information. Inconsistent resources fragment our knowledge of the language. Integrating data would make it possible to give a complete description of the language from the earliest times to the present day.

I am a researcher at the Institute of Polish Language, Polish Academy of Sciences (IJP PAN) which is the leading centre of lexicographic research in Poland. The collection of IJP PAN includes scholarly dictionaries of the Polish language, based on linguistic material collected by several generations of researchers – historical, dialectal, onomastic and present Polish dictionaries. This collection is highly heterogeneous, apart from electronic dictionaries (completed and still ongoing), it includes printed dictionaries, supplements, word card catalogues and corpora.

IJP PAN has collected a huge amount of lexicographic data. As a result, many challenges have arisen, such as how to store this data effectively and how to integrate the separate resources. Together with a team of IJP PAN employees, I am working on a research project called Dariah.lab: „Digital Research Infrastructure for Arts and Humanities" (POIR.04.02.00-00-D006/20; https://lab.dariah.pl/), where we are using existing and original solutions developed for the

digitization, reconciliation and integration of lexicographic data.

The research objective is to develop and adapt methods for processing printed dictionaries (turning printed dictionaries into annotated digital versions), and for integration of lexicographic data. IJP PAN resources, provide a challenge to existing processing and integration methods. Firstly, the nature of the data is varied (highly structured dictionaries of the Polish language and extensive corpora). What is more, textual data is expected to be integrated with resources currently available in the form of images: the word card catalogues, there are thousands of paper and partially digitised cards, store fragments of historical or dialectal uses of the language.

Research Visit

Thanks to the e-Lexis project of visiting grants, I had the opportunity to visit the Instituut voor de Nederlandse Taal (INT) in Netherlands. INT studies all aspects of the Dutch language, including its vocabulary, grammar and linguistic variations. This institute has a long lexicographical tradition and a rich collection of lexicographical resources.

On the first day of my stay, there was a seminar in which I introduced the INT team to the tasks we are carrying out within the Dariah.lab project, related to processing printed dictionaries into a structured digital versions, enriching dictionaries, and integrating lexical resources. My host at INT, Katrien Depuydt, introduced me to the idea of a data-based integrated lexicographic infrastructure, that makes it possible to describe a language across the centuries and in all its complexity. INT is a leading center for e-lexicography. Projects for the integration, coordination, and stimulating the scientific description of language represent a high level of advancement.

Integrating lexical resources

A centralized model of language data management has been developed and implemented at INT to stimulate work on historical and contemporary resources. The INT collects data on grammar, terminology, neologisms, and dialects of the language. The different layers of language description form modules that together provide a complete overview of the language in its dynamics and complexity. The prepared workflow and the way the modules are organized enable the generation of multiple links between the resources.

The model applied at INT assumes a centralised management of resources. Data from dictionaries and corpora power a central lexicographic database. Particular elements of the entries build the layers of the diachronic module, that can be easily implemented in various projects. INT also provides users with an integrated dictionary portal, which makes it possible to search in all

of these dictionaries. A wide set of linking facilities allows users to move freely between resources.

Linking existing resources: the historical dictionary portal

Four historical dictionaries, collected together, describe the Dutch language from about 500 to 1976: Oudnederlands Woordenboek (ONW, Dictionary of Old Dutch, 500-1200); Vroegmiddelnederlands Woordenboek (VMNW, Dictionary of Early Middle Dutch, 1200-1300); Middelnederlandsch Woordenboek (MNW, Dictionary of Middle Dutch, 1250-1550); Woordenboek der Nederlandsche Taal (WNT, Dictionary of the Dutch Language, 1500-1976). They provide the core material for the historical lexicographic infrastructure developed at the INT. The four dictionaries have been converted to a standardize TEI encoding and interlinked at the lemma level.

Modular approach to development of lexical resources

The GiGaNT lexicon has two main modules: GiGaNT Hilex – the historical lexicon component and GiGaNT Molex – modern lexicon component.

Work on modern Dutch is increasingly carried out in a centralized way. The *Algemeen Nederlands Woordenboek* (ANW) is a dictionary of contemporary Dutch, with advanced search functions. The ANW infrastructure connects to the project describing the newest Dutch vocabulary – neologisms. These two together are linked to the MoLex central module.

Conclusion

I came to INT to gain knowledge for processing and integrating lexicographic data from the diachronic perspective. The main research objective is to prepare the concept of integrating data from historical dictionaries created at the IJP PAN. I wanted to find the answers to the questions of how to integrate data from several dictionaries in an automatic way, and how to create the search for links at the level of headwords, word meanings, and inter-word relations to show the language from a diachronic perspective.

In Dariah.lab we are involved in complex lexicographic data processing – starting with OCR, moving on to post-correction and ending with the dictionary segmentation phase. Information in dictionaries is often abbreviated and highly condensed. We recognize the text using OCR tools (Tesseract and AbbyFine Reader). Tesseract, while giving quite good recognition results, does not preserve the typography of the text, which provides a lot of semantic information. During the visit, I was introduced to INT projects for which recognition of text and its structure has been a challenge, e.g. in the Couranten Corpus comprises the seventeenth-century Dutch newspapers

(https://couranten.ivdnt.org/). Tools for automatic detection of regions, lines and words as Tesseract, Ocropy are useful in digitisation work on printed resources. Grobid and Transkribus can be used for layout analysis. The latter can be used both to correct recognized text and provide a better starting point for the semi-automatic structuring of dictionary entries.

Thanks to the e-Lexis project, I could see how a central language data management system was implemented here. INT's works on digitising printed dictionaries and integrating lexical resources are very advanced. The experience gained from this visit is a great help in the project currently ongoing. All ideas and suggestions will be discussed with the project team. The key to integrating dispersed resources is:

1. defining a consistent data model:
    a) identifying the modules (historical, modern, dialectal lexicon components);
    b) describing relevant relations between modules.
2. implementing an identifier system for electronic resources (lemmata, superlemmata, definitions, citations, and sources).
3. the construction of modules for the different layers of language (module for historical lexis, module for modern language, module for dialects), which are first linked within a module, then it is possible to create links also between these modules.

The works carried out on integrating the dispersed resources of IJP PAN are conducted within the Dariah.lab research project. As a result of integration, a unique database of lexicographic resources will be created. Access to the database will be possible via a modern and easy-to-use WWW interface.

Acknowledgements

I would like to express my sincere appreciation to all at INT, to Katrien Depuydt for her wonderful efforts as my host, for sharing with me her vision of modern e-lexicography and for showing me how this vision is being implemented at INT; to Carole Tiberius for preparing this visit and for all the valuable organizational tips; to Thomas Haga and Roland de Bonth for information about historical dictionaries at INT; to Vivien Waszink and Boukje Verheij for introductions to modern language projects; to Katrien and Jesse de Does for valuable advice on converting printed dictionaries into digital versions and for information on available tools; to Veronique de Tier for a presentation on dialect projects and introducing working methods; to Frank Landsbergen for a

presentation on grammar portals; to Dirk Kinable for a presentation on a terminology integration project. I would like to thank the INT team for their valuable support and for sharing their great experience with me.

ELEXIS TRAVEL GRANT REPORT

Report on Elexis Transnational Research Visit Grant at INSTITUUT VOOR DE NEDERLANDSE TAAL, LEIDEN, THE NETHERLANDS

(Leiden, Netherlands, June 13 – June 22, 2022)

Prof. Dr. Carolin Müller-Spitzer

Project title: How is the corpus influencing collocation sets in dictionaries? Enhancing a study on German collocation sets for *Mann* and *Frau* to a contrastive German-Dutch study

The stay financed by the Elexis Travel grant was used to work together with Carole Tiberius and other colleagues at INT on a contrastive German-Dutch study. It is on the influence of the corpus on collocation sets in dictionaries, especially for the entries *man* and *woman*. The initial study on German will be briefly outlined in the following.

Initial Study

The initial study (Müller-Spitzer & Rüdiger 2022, Müller-Spitzer & Lobin 2022) was about the influence of the corpus on collocation sets in dictionaries, exemplified with the entries *Mann* and *Frau.* The starting point of this study was the observation that even in modern corpus-based dictionaries of German, e.g. elexiko[1], the descriptions of entries such as *man* or *woman* are more influenced by stereotypes than we expected. In elexiko, collocation sets are listed for each keyword. In the case of *Mann* (man) and *Frau* (woman), the selection of the most frequent collocates leads to very different representations. It is particularly striking that in the article *Mann*, the agent role constitutes the second collocation set ("What does a man do?"), whereas in the case of *Frau*, the patient role ("What happens to a woman?") is listed second; an imbalance that some researchers have already criticised as *doing gender* (Nübling 2009; Hu, Xu & Hao 2019; Hidalgo Tenorio 2000). The fact that this is presented in the dictionary in this way is due to the frequency of the groups, i.e. in the case of women, the patient role is much more prevalent in the corpus texts of the elexiko corpus than the

[1] https://www.owid.de/docs/elex/start.jsp.

agent role. For men, it is the other way round. Another example of stereotypical representation of gender roles is Duden Online[2], but only a certain item class, namely the computer-generated collocation profiles. Typical adjectives for *Mann* are *young*, *old*, *rich*, *strong*, *adult*, *powerful*, *armed* and *right*, whereas the typical ones for *Frau* are *young*, *old*, *beautiful*, *tall*, *naked*, *pregnant*, *gracious* and *employed*.

The corpora on which the two dictionaries (elexiko, Duden Online) are based are - like the large linguistic corpora on German in general - dominated by newspaper texts (critical to unbalanced corpora in the context of lexicography cf. Rundell & Atkins 2013: 1339). In our case study for German, we show how the linguistic contexts of *man* and *woman* obtained on the basis of newspaper texts differ from other samples, e.g. texts of fiction or popular magazines, and how different the 'reality' shown in the dictionary would look if the corpus was composed differently.

One example are the verbal co-occurrences for *Mann* (cf. Fig. 1), i.e. fillers to collocation sets like "What does a *man* do?" or "What happens to a *man*?". Verbs in the fictional books are *scrutinize, marry, observe, sit opposite, turn to (mustern, heiraten, beobachten, gegenübersitzen, zuwenden*). In magazines, words referring to love life, money or power are frequent collocators: *marry, fall in love, question, earn, cheat,* or *dominate* (*heiraten, verlieben, befragen, verdienen, betrügen, , dominieren)*. In the newspaper texts, the context of violence is predominant: *arrest, assault, threaten, shoot,* and *rape* (*festnehmen, überfallen, bedrohen, erschießen, vergewaltigen*) are particularly significant co-occurrences. Accordingly, the 'linguistic reality' differs greatly depending on which corpus we chose to analyse collocations (and so would do the lexicographic entries).

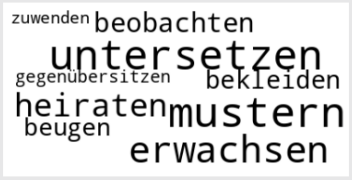
VV	fiction books	magazines	newspapers
Mann			

Fig.1: Verbal co-occurrences for *Mann* (the font size depends on the significance based on Poisson-distribution).

Our results for German show that newspaper texts preferably display differences between men and women instead of making common features, characteristics and actions the subject of discussion. The context of violence, for example, which is particularly over-represented in the elexiko entries,[3] is dominant only in the newspaper corpus. It becomes clear that the

[2] http://www.duden.de.

[3] In the entry *man* in elexiko, the first three verbal co-occurrences are *dominate, murder* and *shoot*.

corpus basis can bring an unnecessarily strong bias towards *doing gender* into the dictionary (cf. also Nübling 2010: 620). The aim of the joint study during the guest stay at the INT was to explore whether the same bias (newspaper-heavy corpora = context of violence) can also be observed in Dutch corpus collections.

Joint Work in at the INT in Leiden

The stay in Leiden was used to cooperate (with Carole Tiberius and other colleagues there, cf. e.g. Steurs et al. 2021) to enhance the study described above with data from Dutch. The INT was chosen because it has the corpora, the corpus analysis systems and the knowledge necessary for the analyses.

We needed the first time of the guest stay to find comparable corpus resources. First, we took the corpus the Woordcombinaties[4] project uses. It contains contemporary language material that mainly comes from newspapers (NRC and De Standaard, thus texts from the Netherlands and Flanders, from 2012-2018) and consists of just over 230 million words. We took this corpus as a 'newspaper-corpus'. Second, we chose the ANW literature corpus, which is a subcorpus of the ANW corpus[5]. The Corpus of Literary Texts contains essays, novels, stories and drama, both original and translated work. The selection takes into account a balanced spread in time and a reasonable distribution between North (Netherlands) and South (Belgium). It consists of approx. 20 million tokens.

The analyses of the different collocation sets show strikingly similar results as the study for German. In fiction texts the verbal collocates for *vrouw* and *man* are very similar to each other; in the newspaper texts, they differ greatly, and are especially affected by the discourse around violence (cf. Fig 2 & 3).

It is also striking how similar the collocation sets are regarding individual lexical items (German/Dutch, cf. Fig 4). In both languages, we find highly significant verbal collocates like *to arrest* (*festnehmen/arresteren, oppakken, annhouden*), *to judge* (*veroordelen/verurteilen*), *to shoot* (*doodschieten/erschießen*).

[4] https://woordcombinaties.ivdnt.org/.
[5] https://anw.ivdnt.org/anwcorpus.

vrouw	man
	

Fig.2: Verbal co-occurrences for *vrouw* and *man* in Dutch fiction texts.

vrouw	man
	

Fig.3: Verbal co-occurrences for *vrouw* and *man* in Dutch newspaper texts.

Mann	man
	

Fig.4: Verbal co-occurrences for *Mann* and *man* in German and Dutch newspaper texts.

The Woordcombinaties team at INT using the corpus which is dominated by newspaper texts also analysed the lexemes *man* and *vrouw* during the guest stay. Here, too, it can be seen, especially in the 'object-of-relation', that many of the verbal collocates come from the context of violence (n = 15/39; n = 5/15 as victim, n = 10/15 as offender). For comparison: 9 of 14 verbs in elexiko listed as collocation fillers for "What does a man do?" can be assigned to the context of violence (*dominieren, ermorden, erschießen, schießen, (sich) verletzen, sterben, stürzen, töten, vergewaltigen*). We proofed additionally whether the comparison between texts from the Netherlands and texts from Belgium show any major differences. However, this is not the case (cf. Fig. 5). Thus, the contrastive results strongly suggest that it seems to be indeed a newspaper-bias, not a specific feature of the German-language corpora we used for our initial study.

Fig.5: Verbal co-occurrences for *Mann* and *man* in German and Dutch newspaper texts.

It was very interesting to discuss these results with colleagues at INT, both one-on-one and in connection with a talk I gave there, since the results raise several questions, e.g.: i) regarding corpus selection: How suitable are newspaper-heavy corpora as a basis for lexicographic analysis when the context of violence is so particularly present there? ii) Do lexicographers have to intervene when the corpus brings a strong *doing gender* into the dictionary? A good compromise seems to be first to research language use with as much reflection (and self-reflection) as possible, and then - as a lexicographer does with offensive or vulgar expressions - to find a balance between language use orientation and the handing-down of outdated role models.

Perspectives

The analyses seem promising enough for us to continue our joint work and probably submit a paper together for the next eLex conference. However, there are still some details to be clarified, especially regarding the standardisation of the collocation measures and the

comparability of the corpora. Without the cooperation with the INT as a starting point, such a joint study would not have been possible.

Bibliography

Hidalgo Tenorio, Encarnación. 2000. Gender, Sex and Stereotyping in the Collins COBUILD English Language Dictionary. *Australian Journal of Linguistics*. Routledge 20(2). 211–230. https://doi.org/10.1080/07268600020006076.

Hu, Huilian, Hai Xu & Junjie Hao. 2019. An SFL approach to gender ideology in the sentence examples in the Contemporary Chinese Dictionary. *Lingua* 220. 17–30. https://doi.org/10.1016/j.lingua.2018.12.004.

Müller-Spitzer, Carolin & Rüdiger, Jan-Oliver 2022. The influence of the corpus on the representation of gender stereotypes in the dictionary. A case study of corpus-based dictionaries of German. Euralex Proceedings 2022 (forthcoming).

Müller-Spitzer, Carolin & Lobin, Henning. 2022. Leben, lieben, leiden: Geschlechterstereotype in Wörterbüchern, Einfluss der Korpusgrundlage und Abbild der sprachlichen ‚Wirklichkeit'. In Gabriele Diewald/Damaris Nübling (eds.) *Genus, Sexus, Gender - Neue Forschungen und empirische Studien zu Geschlecht im Deutschen* (Reihe Linguistik: Impulse und Tendenzen), S. 35–64.

Nübling, Damaris. 2009. Zur lexikografischen Inszenierung von Geschlecht. Ein Streifzug durch die Einträge von Frau und Mann in neueren Wörterbüchern. De Gruyter 37(3). 593–633. https://doi.org/10.1515/ZGL.2009.037.

Rundell, Michael & Beryl T. Sue Atkins. 2013. Criteria for the design of corpora for monolingual lexicography. In Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard & Herber Ernst Wiegand (eds.), *Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography*, 1336–1343. De Gruyter Mouton. https://doi.org/10.1515/9783110238136.1336.

Steurs, Frieda, Kris Heylen & Vincent Vandeghinste (2021). Hoe automatische vertaling de gender bias van AI verraadt. In: Wat gebeurt er in het Nederlands?. Sterck De Vreese. ISBN: 9789056158033.

Annika Simonsen, linguist
Talutøkni, the Faroe Islands

Report on Elexis Transnational Research Visit Grant at Det Danske Sprog- og Litteraturselskab (DSL, Denmark) & University of Copenhagen (UCPH, Denmark)

(Copenhagen, Denmark, November 16th – 27th, 2021)

I was given the opportunity to visit Det Danske Sprog- og Litteraturselskab (DSL, Denmark) and the Centre for Language Technology (CST, Denmark) at the University of Copenhagen (UCPH, Denmark), who together with Dansk Sprognævn (DSN, Denmark) form one of the most important research centres for NLP in the North, specialising in language technology for the West Nordic languages. During my visit, I met with several experts, who shared their valuable knowledge with me. Aside from getting assistance with my own work, I made several contacts at both DSL and CST, and opened up the possibility for future collaboration.

My project, Project Ravnur, is a Faroese speech recognition project that creates versatile language materials for a broad range of Faroese LT (a so-called BLARK, or Basic Language Resource Kit). My current role as linguist in our project is to oversee our transcription assistants, but I am also creating a wide-coverage dictionary including phonetic information for all word forms and Part of Speech (PoS) tags. My research visit was originally planned to be in fall of 2020. My task at the time was to prepare a PAROLE tag definition for Faroese and apply the tagset in the dictionary and transcription corpora. I applied to visit DSL and CST at UCPH because they are the Danish partners in the PAROLE-project; this made them my ideal hosts.
My visit was pushed back a year and by that time I had already completed my PAROLE tag definition and had moved onto other things. However, I still had two things I wanted to learn from DSL and CST:

1) Det Centrale Ordregister (COR)
I am currently working on a GUID-style index system called OTAL (O-number) to use for all linguistic data in Project Ravnur. The index system is developed for implementing the system into the new Faroese orthographic dictionary, when it has been completed. The index system takes its main inspiration from the ongoing Danish project, COR (the Central Danish Word Register). COR is a collaboration between DSN, DSL and CST. It is an ongoing project, which means I would get to witness the work in progress. Furthermore, I could get assistance with the Faroese OTAL.

2) General insight and contacts
My second goal that I had in mind was to get general insight into as many relevant projects at DSL and CST as possible. Both institutions are home to some of the leading researchers in the North. Being able to shadow their work would be a valuable opportunity, and I would strive to create contacts for Project Ravnur.

Det Danske Sprog-og Litteraturselskab (DSL)
At DSL I was introduced to several dictionary editors and their work. I was given the opportunity to explore DSL's tools and resources on my own. My first meeting was with a senior editor, who introduced me to DSL's corpus tool, CoREST, and the PAROLE tags that are used in it. I learned how DSL obtains their text for the corpus and how their metadata is stored. I am interested in corpora and text collection, because Faroese lacks a big text corpus and Project Ravnur is currently in the process of creating one for our language model. It was interesting to see how CoREST has been created and to learn about the potential challenges that come with maintaining a big text corpus.
Other relevant resources that I was given more insight into was Den Danske Ordbog or DDO (an online dictionary of Modern Danish) and its structure, as well as the Den Danske Begrebsordbog (the Danish thesaurus), which categorises words according to their relatedness - something that is integrated in DDO as well as in other projects.

Not only did I learn about DSL's own projects during my visit, but I was also given the opportunity to show them my own work and ask for advice. It is not often that I get to ask for advice from someone who has experience both with lexicography and language technology like the editors at DSL do. The feedback I received on our ongoing OTAL project was very helpful. Furthermore, now I have made contacts who are familiar with my project and whom I can reach out to next time I have any questions.

The Centre for Language Technology (CST) at the University of Copenhagen (UCPH)
At CST I was introduced to the research staff and their work. It would not be possible for me to mention every single thing I learned at CST, so I will only include the highlights. Among the many relevant resources I learned about was DanNet, a wordnet for Danish. DanNet shows how concepts relate to other concepts, e.g. that cake (*kage*) is used for eating (*spise*) and is made of sugar (*sukker*) and flour (*mel*).

Some researchers at CST even showed an interest in developing tools for Faroese. I provided them with the necessary materials and they trained a lemmatizer (CSTlemma) on Faroese. As of writing, we are testing several tools for Faroese and we are working together to improve them. This was an unexpected, but welcomed opportunity for me, and I am delighted to have Faroese LT tools in the making. Furthermore, a research assistant at CST generously donated her time to program scripts for me to use to collect Faroese news text from the internet. These scripts are making it possible for us in Project Ravnur to proficiently collect text for our background text corpus, which we will use for our language model.

Other activites
I was lucky to part-take in several extra activities during my research visit. At the very beginning of my visit I got to present a poster about Project Ravnur and OTAL at the Language Technology conference (Sprogteknologisk Konference) at CST. Later I attended professorial inauguration lectures in Danish language at UCPH, as well as a virtual European Language Resource Coordination (ELRC) conference. I also managed to take a day-trip to Bogense to visit Dansk Sprognævn (DSN). And finally, on one of my last days, I was given a tour of historical Faroese documents at the Arnamagnæan Manuscript Collection, which were being restored.

Conclusion

As someone working with an extremely low resource language such as Faroese, it was helpful to visit institutions who also work with a low resource language such as Danish, because I was able to learn from experienced people who have faced similar challenges as myself. There is no established Language Technology (LT) field in the Faroe Islands as of yet, but it is easy to recognize the need for one. With the arrival of the digital age, Faroese could be facing digital extinction if it cannot keep up. Developing good LT resources for Faroese will therefore play a crucial role for the long-term survival of Faroese. The things I have learned at DSL and CST are going to guide me when I create resources for Faroese, because I have seen how the Danish resources have been made and I have seen the challenges they have faced and how they solved their problems. On top of having learned a lot, I now know several familiar faces at DSL and CST, who I can contact next time I need guidance. I am planning to stay in touch.

I would like to thank Bolette Sandford Pedersen and Sussi Olsen at CST and Sanni Nimb at DSL for their wonderful efforts while being my hosts, and also every expert at DSL and CST who shared their knowledge and experience with me, especially Thomas Troelsgård at DSN for helping me with OTAL; Bart Jongejan and Nathalie Hau Sørensen at CST for creating LT tools for Faroese with me; and Peter Juel Henrichsen at DSN for encouraging me to apply for the ELEXIS grant.

Dóra Mária Tamás
Senior Terminologist at the Hungarian Office for Translation and Attestation Ltd. and
Lecturer at the Budapest University of Technology and Economics
(tamas.dora.maria@gmail.com)

Report on the Elexis Transnational Visit Grant at the Dutch Language Institute (Instituut voor de Nederlandse Taal)

(Leiden, Netherlands, 13 June – 17 June 2022)

Project title: The Description of the Characteristics of Legal Terminology and Issues of Editing Legal Databases in Lexicography and Terminology

1 Introduction

I applied for a visit to Dutch Language Institute with the following goals in mind. The project proposal that I submitted to the call for ELEXIS grants for research visits included the plan to discuss and collect sources about the features of legal LSP and terminology, and to get familiar with the upcoming project of an electronical legal dictionary of the Institute.
Since I am working on my habilitation book, I was interested to get an insight into the research activities and experiences of the local experts. I was motivated to choose the Dutch Language Institute, because I was highly impressed by lectures held by Professor Frieda Steurs at an international conference in this field of my interest.

2 Aims of the Project

The goals of my visit were to gain knowledge and valuable experience through discussions in personal meetings with experts of the field and the collection of new relevant information and sources for my research and book editing in both areas, namely the features of legal terminology and the editing of legal electronical dictionary at the Institute.

3 Description of the Research Visit

My Research Visit has been prepared thoroughly with a valuable programme containing prescheduled meetings. On the first day of my visit, I was introduced to the staff and the main activities of the Dutch Language Institute. On the second day I had the opportunity to attend a meeting with Ms. Theresa Munneke, the editor of Juridisch Woordenboek, the upcoming legal database containing the terms of the Dutch legal system, who explained in detail the editing principles, working methods, structure, content, classification, selection of sources and challenges, and included in her presentation even some specific examples. I had the possibility to pose questions and discuss all the interesting issues related to the digitalisation and editing of the legal dictionary. I have received the link to study the online version.

Following the interesting meeting, I had another worthy personal discussion organised with Mr. Tony Foster (Lecturer at the Leiden University Centre for Linguistics, University Leiden), with whom we shared our views and experiences in the field of legal terminology and translations. It was interesting to learn about the difficulties and solutions from the point of view and for the Dutch legal language and terminology. The exchange of ideas helped me to strengthen and broaden my knowledge.

On the same day I gave a presentation about my personal experiences, i.e. studies, work, research and teaching activities, and my ongoing project. I have been working for 15 years at the Hungarian Office for Translation and Attestation Ltd., where first I was a translator of certified translations and court interpreter, and after I acquired my PhD in Terminology at ELTE University, I became a terminologist at the office, where I am currently responsible for IUSTerm, an internal termbase for terms of law and public administration as a Senior Terminologist. I have also experience in teaching at university courses, which I started in 2013. The presentation was followed by an interesting discussion, which provided useful feedback for my project. After receiving precious information about useful sources, I was provided of an access to the university library. On the third day during my stay, I had the opportunity to attend an interesting workshop titled ‘Is there a place for Google Translate in higher education?’ organised by the University of Leiden with a few presentations in English. After the symposium I spent my time at the University library. On the fourth day I met a colleague of the institute to discuss the characteristics and challenges of working out the terminology of education, which only at first glance seems to be a clary systemised terminology. We exchanged our experiences in this field. I had the possibility to attend the interesting presentation of Ms. Caroline Müller-Spitzer, a lexicographer from Germany, who was also a winner of the ELEXIS grant. I spent the rest of my time by searching for valuable sources at the University library. This was in line with my plan to have to possibility for individual research after receiving a brief guidance and

assistance: searching for the relevant literature in order to prepare a habilitation work about terminology with a focus on legal texts and termbases.

4 Concluding remarks

The visit has been useful and informative in all the aspects mentioned above. It has been very fruitful to me to have an opportunity to receive guidance, advice and consultations from the experts in the field of lexicography and terminology and other specialists from 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 one week. I can say that it succeeded beyond my expectations, and I believe that the experience and contacts I have gained from it will be worthy for my future work.

5 Acknowledgments

I would like to use this occasion to thank my hosts – all the staff of the Dutch Language Institute – for their hospitality and for making me feel at home during my stay in Leiden. Furthermore, I would like to thank all the editors and researchers for eagerly answering all of my questions and providing me with additional material on the topics I found interesting. In particular, I would like to express especially my thanks to Professor Frieda Steurs for her wise pieces of advice and interest shown in my work and research, and Ms. Carole Tiberius for assisting me prior to and during my visit as well as planning my activities. I very much trust that we will be able to enhance our cooperation and common work in the future. Finally, I am grateful to the ELEXIS project for making my research visit possible.