elexis european lexicographic infrastructure

D.9.1 Report on trans-national access – year 1

Author(s): Bolette S. Pedersen,

Sussi Olsen

Date: 31-07-2019

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

NUMBER AND NAME OF THE DELIVERABLE

Deliverable Number:	9.1
Dissemination Level:	PU
Delivery Date:	31-07-2019
Version:	V0.1
Author(s):	Bolette S. Pedersen Sussi Olsen

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 03/06 2019	First draft	Bolette S. Pedersen, Sussi Olsen
	First draft and review	Iztok Kosem
V1.0 01/07 2019	Final version	Bolette S. Pedersen, Sussi Olsen

Report on trans-national access Year 1

Table of Contents

1 Introduction: Trans-national access – Year 1

This deliverable presents the results of the two first calls and grant visit rounds of the transnational access program of the ELEXIS project. The first call was launched in summer 2018 when ELEXIS had been running for approximately half a year, and the second call was launched in winter 2018. Overall, 10 visits have been granted during the first year, and at the time of writing all five visits of the first round have been successfully completed.

In the following sections we present the objectives of the visits and describe the procedures we have established in order to run the calls and visits as smoothly as possible. Secondly, we provide the scientific reports from each of the five grant holders of the first round; each of them documents the outcomes of each visit and their relation to the particular project of the grant holder. Finally, we provide some overall comments regarding the benefits for ELEXIS of the visit program in year one.

2 Objectives

Even though lexicography has a long history of international research conferences, it has traditionally been a research area with limited knowledge exchange outside of each lexicographical institution, and in many cases lexicographic data has only been accessible to researchers from the institution who created the data and held the copyright.

This tradition is partly connected to the fact that practical lexicography has a strong commercial basis; lexicographical data used to be good business. But it also relates to the fact that enabling easy access to restricted data requires significant effort into facilitating and controlling this access - which again requires time and money not easily found in the budgets of lexicographic projects.

To this end, an important objective of ELEXIS is to stimulate knowledge exchange between lexicographical research facilities, infrastructures and resources throughout Europe, which can consequently mutually benefit from the vast experience and expertise that exist in the community.

Inspired by other EU projects such as EHRI[1], RISIS[2], InGRID[3], and sobigdata[4], ELEXIS offers trans-national access activities in the form of visiting grants that enable researchers, research groups and lexicographers to work with lexicographical data which are not fully accessible online. Furthermore, grants offer access to professional on the spot expertise in order to ensure and optimise mutual knowledge exchange. Finally, travel grant recipients can gain

[1] https://ehri-project.eu/ehri-fellowship-call-2016-2018

[2] http://datasets.risis.eu/

[3] http://www.inclusivegrowth.eu/visiting-grants

[4] http://www.sobigdata.eu/access/transnational

knowledge and expertise by working with lexicographers and experts in NLP and artificial intelligence.

The trans-national access activities are expected to have a long-term impact specifically but not only for lesser-resourced languages, boost the network and infrastructure of the European lexicographic community, and facilitate future collaboration and knowledge exchange.

The objectives of the ELEXIS trans-national activities can be summarised as follows:

- 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.

The trans-national activities represent a way of ELEXIS to enable access to restricted data, which has so far not been available outside of the hosting institutions, to researchers from other institutions and countries. And as the results of research conducted in trans-national activities will be available under open-access licences, the international lexicographic community will become acquainted with previously inaccessible resources.

3 Application procedure

The visits

The transnational activities consist of visiting grants of 1 to 3 weeks for researchers to experiment with and work on lexicographical data in a context of mutual knowledge exchange with the hosting institutions. Five visiting grants are made available twice a year during the entire project period, amounting to 35-40 grants in total. Researchers and lexicographers within the EU member states and associated countries have been invited to apply for a visit of free access to and support from one of the lexicographical institutions.

The infrastructures/lexicographical institutions

The following 11 lexicographical institutions accept research visits during the ELEXIS project:

- ELEXIS-SL: Institut Jozef Stefan (JSI, Slovenia)
- ELEXIS-NL: Institute for Dutch Language (INT, The Netherlands)

- ELEXIS-AT: Austrian Academy of Sciences (OEAW, Austria)
- ELEXIS-RS: Belgrade Center for Digital Humanities (BCDH, Serbia)
- ELEXIS-BG: Institute of Bulgarian Lanugage Lyubomir Andreychin (IBL, Bulgaria)
- ELEXIS-HU: Hungarian Academy of Sciences (RILMTA, Hungary)
- ELEXIS-IL: K-Dictionaries (KD, Israel)
- ELEXIS-DK: Det Danske Sprog- og Litteraturselskab, University of Copenhagen (DSL/UCPH, Denmark)
- ELEXIS-DE: Trier Center for Digital Humanities (TCDH, Germany)
- ELEXIS-EE: Institute for Estonian Language (EKI, Estonia)
- ELEXIS-ES: Real Academia Española (RAE, Spain)

During the visits, the hosting institutions provide support in terms of lexicographical and IT expertise.

The calls

The calls for applications include descriptions of the institutions and the lexicographical resources, tools, and expertise that are made available. Researchers and lexicographers interested in visiting a particular host institution are encouraged to make motivated applications describing their background, the purpose of the visit etc.

Applicants are also encouraged to contact the institution of interest before applying in order to dis cuss the suitability of their proposal.

The first call was launched on June 6th 2018 with a deadline for applications August 6th. The second call was launched December 13 2018 with a deadline January 13 2019.

For illustration, the first call had the following wording:

> ***Five visiting grants for transnational access to ELEXIS Infrastructures: Apply for visiting grants here!***
>
> *ELEXIS is a European Infrastructure which aims to foster cooperation and knowledge exchange between different research communities in lexicography in order to bridge the gap between lesser-resourced languages and those with advanced e-lexicographic experience.*
>
> *A key activity of the project is to provide transnational access to existing lexicographical milieus in Europe including the resources, tools and expertise of these groups.*
>
> *Via visiting grants (of 1, 2, or 3 weeks), researchers will be able to experiment with and work on data in a context of mutual knowledge exchange.*
>
> *Researchers are invited to apply for free-of-charge access to and support in one of the 11 ELEXIS infrastructures (links to infrastructures).*
>
> *First deadline for applying is August 6 covering visits for the period October 2018 to April 2019. New calls will be launched every six months.*
>
> *Application form (link)*
>
> *Terms and conditions for applying for a visiting grant (link)*
>
> *Review procedure of application: The ELEXIS TA Review Committee will review the applications and notify the applicants within approximately a month of whether the grant has been obtained.*

Principles for reimbursement (link)

Dissemination of calls

The calls are disseminated through the ELEXIS website https://elex.is/grants-for-research-visits/, mailing lists and newsletters, as well as via Facebook and Twitter.

The Reviewing Process

For each call, the Transnational Access Review Committee reviews and prioritises the received applications.

The Transnational Access Review Committee has been appointed by the Technical Management Board and consists of the following partners:

- Professor Bolette S. Pedersen, University of Copenhagen
- Dr John McCrae, National University of Ireland, Galway
- Dr Carole Tiberius, Instituut voor de Nederlandse taal

Before notifying the applicants, the committee contacts the hosting institutions to ensure that the proposed projects are in accordance with the hosting institution's field of activities and that the hosting institution can offer assistance to the project.

4 Status year 1

The first two calls each received 10 applications, five of which were selected (50% acceptance rate).

As part of the first call, two grant holders visited Spain, while the other three went to Germany, the Netherlands and Israel respectively.

Home institution	**Hosting institution**	**Project**
Austrian Centre for Digital Humanities (Austrian Academy of Sciences), Austria	ELEXIS-ES: 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	ELEXIS-ES: 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	ELEXIS-DE: 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	ELEXIS-NL: Dutch Language Institute (INT, The Netherlands)	A multisemiotic e-dictionary
Copenhagen Business School, Denmark	ELEXIS-IL: K Dictionaries (KD, Israel)	Business Model Innovation in the Dictionary Industry

As part of the second call, two grant holders are going to visit the Netherlands, while the rest will go to Bulgaria, Slovenia, and Denmark respectively.

Home institution	**Hosting institution**	**Project**
Sofia University "St. Kliment Ohridski", Bulgaria	ELEXIS-RS: Belgrade Center for Digital Humanities (BCDH, Serbia)	Encoding Latin-Bulgarian Dictionary
Institute of Polish Language Polish Academy of Sciences, Poland	ELEXIS-NL: Dutch Language Institute (INT, The Netherlands)	Integration of four Polish old and modern dictionaries
CELGA-ILTEC, University of Coimbra, Portugal	ELEXIS-SI: Institut Jožef Stefan (JSI, Slovenia)	Improving a procedure for automatic extraction of data and import into DWS
Institute of Croatian Language and Linguistics, Croatia	ELEXIS-DK: 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	ELEXIS-NL: Dutch Language Institute (INT, The Netherlands)	English - Georgian Maritime Dictionary

Higher popularity of certain hosting institutions means that these institutions will be unavailable for further visits due to the fixed budget for travel grants given to each hosting institution. For example, the Dutch Language Institute has already hosted three visits and will thus not be included in the third call.

5 Scientific reports from the first round

Each of the five grant holders from the first call has written a report about their research visit. The reports are published on the ELEXIS website, https://elex.is/travel-grant-reports-call-1/.

The reports are inserted below.

Asil Çetin (asil.cetin@oeaw.ac.at)
Austrian Centre for Digital Humanities at Austrian Academy of Sciences
Vienna, Austria

Report on Elexis Transnational Research Visit Grant at Real Academia Española

(Madrid, Spain, April 7 – April 27, 2019)

Project Title

Visual Data Analysis for Multi-Dimensional Corpus Exploration

About the Project

The aim of this project is to develop an explorative data visualization application to analyze and visualize regional language varieties and statistical differences in lexical uses of languages. This project runs as a collaborative design study as part of my Master's thesis at the Computer Science Faculty of the University of Vienna and Austrian Centre for Digital Humanities. The software architecture of the application follows a decoupled service pattern separating data collection / curation, corpus access and frontend of the web application. Hence, a software architecture like that would offer the opportunity to reuse the application for different language sources and different corpus engines.

About the Data Sources

During the design, development and evaluation phases of this project the following two main data sources consisting of large corpora in two different languages are being used:

- **Austrian Media Corpus:** AMC was created as part of a cooperation between the Austrian Academy of Sciences and the Austrian Press Agency. It covers the entire Austrian media landscape of the past two decades, containing 40 million texts, constituting more than 10 billion tokens. AMC ranks among the largest collections of its kind as a contemporary German language corpora.

- **Real Academia Española:** The "Advanced Search Interface" of DLE 23, CORPES, CREA and CORDE are some of the query mechanisms of the Real Academia Española (RAE), which provide accurate linguistic data about varieties of Spanish language. RAE, with its affiliations in 22 hispanophone nations, offers the most extensive knowledge and data regarding the Spanish language.

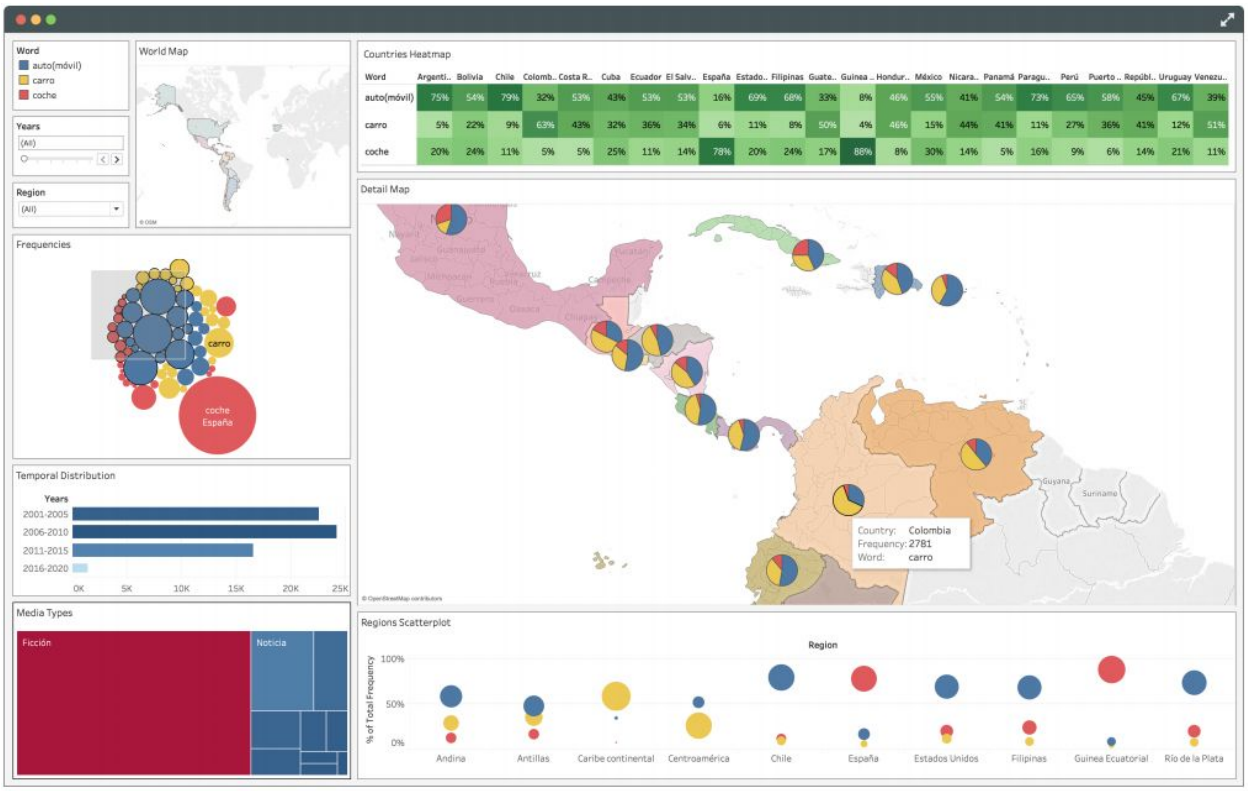


Figure 1: High-fidelity prototype using sample data from Corpus del Español del Siglo XXI of Real Academia Española. The application is intended be used with other languages / corpora as well.

Summary of the Research Visit

My research visit at the Real Academia Española through the ELEXIS travel grant started by the second week of April 2019 and lasted three weeks. In this timespan the main aim of my project was to develop and design an explorative data visualization application to analyze and visualize corpus linguistic data.

Since my project runs as a collaborative design study and focuses on the target group of researchers of the fields of linguistics and humanities, it's crucial to be in contact with domain experts of these fields.

During my research visit at the Real Academia Española it was possible to fulfill both of these expectations for my project: accessing some of the largest and well annotated corpora available today and exchanging knowledge and feedback with domain experts during various stages of my research.

I've worked during the weekdays at the offices of Real Academia Española in Madrid's El Viso quarter, which are allocated for departments of computational and corpus linguistics, lexicography and software development. My main supervisor was Mr. Jordi Porta-Zamorano, PhD and it was great opportunity to work with him on a daily basis and profit from his decades long experience in this field and at Real Academia Española.

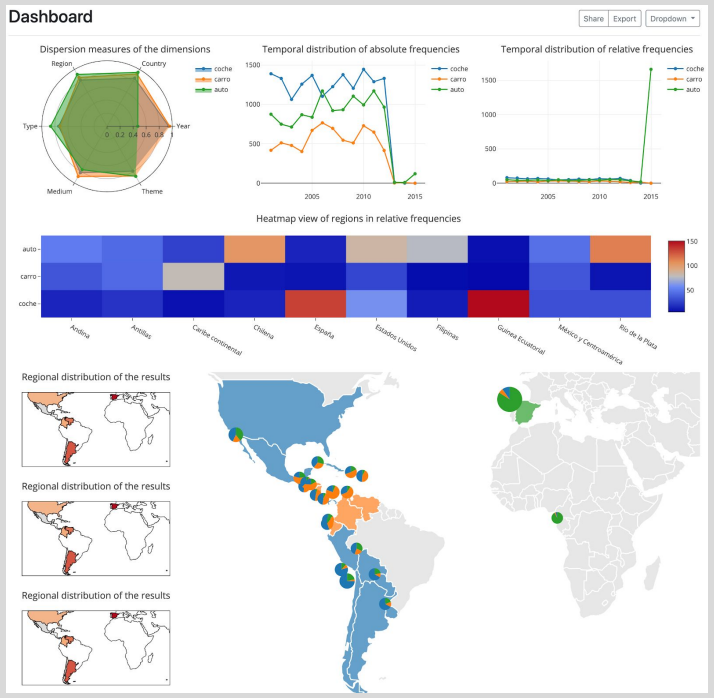


Figure 2: The tool being implemented as an interactive web-app with real-time connection to RAE's CORPES API.

Some of the other colleagues who I've worked with and exchanged knowledge on daily basis were José Luis Sancho, MSc, Rafael Ureña, MSc and Juan Romeu, PhD. During and after each iteration of design and implementation it was possible to consult with these colleagues to review the strengths and weaknesses of various choices.

At the end of my research visit I had the opportunity to make a final presentation about my project with the attendance of ~20 colleagues, which made it possible to gather feedback from various domain experts.

Results and Future Work

The results of this highly productive research visit in terms of software development, design study and knowledge exchange can be listed as follows:

- Accessing large and well annotated linguistic data
- Developing the software tool using the real-time connection the available resources
- Getting valuable input from domain experts on a daily basis during design / development

- Presentation and live-demonstration session with feedback from colleagues at RAE
- Setting up a basis for future collaboration regarding this project with RAE

Currently the results in terms of software tool are still in progress and the public git-repository of the web-application can be found at: https://github.com/asilcetin/corpsum. In the upcoming months the application will be under further development and the results of both development and the design study will be documented in a paper. This paper and detailed documentation about the tool will be made available / linked at main readme page of the git repository.

ELEXIS Transnational Research Visit Grant

Final report – Ana Salgado, NOVA CLUNL

Travel Grant: Call 1
Researcher: Ana Salgado
Hosting institution: ELEXIS-ES: Real Academia Española (RAE, Spain), Centro de Estudios de la RAE y de la Asociación de Academias de la Lengua Española
Institutional affiliation: Faculdade de Ciências Sociais e Humanas da Universidade NOVA de Lisboa/Centro de Linguística da Universidade NOVA de Lisboa, Portugal
Current position: PhD Student in Terminology
Project title: Comparison of the macro/microstructure of the Iberian Academy Dictionaries: the *Diccionario de la Lengua Española* (DLE, RAE) and the *Dicionário da Língua portuguesa* (DLP, ACL)
Period of stay: November 11-30, 2018

Introduction

Since I am the person in charge of the coordination of the new Portuguese Academy Dictionary, specifically as regards planning the macrostructure and microstructure of this dictionary, it is crucial that I learn how other structures are devised, such as the *Diccionario de la lengua española* (DLE) published by an analogous academy – *Real Academia Española*.

This research is aimed at defining guidelines for the inclusion and description of terms in general language dictionaries in order to help lexicographers in this specific task. Combining lexicographical and terminological methods is definitely a plus for the planning of a dictionary's macrostructure and microstructure, improving the organization and description of lexicographical articles as a whole.

RAE

ACL

Research goals

Our main goal is to analyse the macro/microstructure of the DLE (RAE) and compare it with the Portuguese Dictionary (DLP) of the *Academia das Ciências de Lisboa* (ACL), especially to rethink the theoretical and methodological methods of the lexicographical tradition, such as the lexicographical labels that identify specialized lexicon.

Since my PhD project focuses on the lexicographical work methodology, there is great interest in getting to know and understand the methodology adopted by the team working at the *Instituto de Lexicografía* (ILex) of the RAE.

In addition to the methodology itself, it would also be highly relevant to exchange views on how scientific terminology is dealt with, discuss different approaches, exchange impressions on lexicographical works, get to know its infrastructures and computational linguistic tools, and share knowledge.

Placing the structures of two dictionaries – RAE vs. ACL – side by side and comparing them will aid in the decision-making process, namely by allowing us to:

- combine lexicographical/terminological methodologies;
- create a methodology for the selection, description/definition of linguistic and conceptual information (ISO TC37);
- improve the macro- and microstructure of lexicographical works;
- increase the quality of lexical databases.

Methodological plan

- Establishing preparatory contacts with the host organisation
- Preparing an interview guide to become familiarized with the departments: ILex, Corpes, Tecnología (Desarollo, Lingüística Computacional)
- Meetings with experts
- Scheduling research interviews
- Analysing interviews
- Analysing infrastructures
- Working on a final presentation on the Portuguese Academy Dictionary

Research questions

- How do lexicographers include terminological units (terms) in Iberian Academy Dictionaries?
- What is the percentage of terms in Iberian Academy Dictionaries?
- What do domain labels tell the user? Do they indicate a technical word? Are they useful?
- Which terms can you find in Iberian Academy Dictionaries? Are they all marked?
- What domains have already been labelled?
- Is it possible to standardize those labels?
- How can you organize conceptual information?
- What standards should be used? (TEI, Oxygen)

First impressions

As our main goal was to get acquainted with the whole structure that supports the DLE, I only focused on the platforms and tools directly related to the DLE working process, although, of course, there are other features that have a direct influence on the dictionary. I spent a few days exploring the Entorno de Redacción Integrado (ERI), a software platform in JAVA and XML, in which lexicographical work is developed. Regarding the lexicographical work itself, I learned how

the dictionary is updated (documentation, academic proposals, commissions, external queries…). From Enclave, a new service platform, I would like to highlight one of its modules: the *Diccionario avanzado*. The DLE contains a lot of information hidden in its database because of print edition limitations. However, this module now enables combined searches and access the list of words related to a specific domain – e.g., when searching for the word *fútbol*, I was able to find all the definitions in which that word appears.

First results

Having access to the DLE's underlying structure made it possible for us to exhaustively survey domain labels in order to determine the number of domains represented, those shared between the two dictionaries, the ones that differ, and which domains are the most frequent.

This study allowed us to analyse, describe, and compare domain labelling in general language dictionaries, which points to specialized lexicon in those general language dictionaries. We have examined whether label application is systematic or there is an imbalance, and whether there are recent and relevant domains that do not appear. We verified the criteria for differential selection and treatment of terms, which differ between the dictionaries under review. Finally, we looked at other lexicographical resources used in the definitions as domain indications.

Stemming from this work, in particular from the topic of domain labelling, we are pleased to announce the acceptance of a submitted paper (peer reviewed) to the *III Jornadas internacionales sobre investigaciones lexicográficas y lexicológicas* (inLÉXICO2019), which will be held at the *Universidad de Jaén*, on April 4-5 2019. (Salgado, A. & Costa, R., *Marcas temáticas nos dicionários académicos ibéricos: estudo comparativo*).

Conclusions and future work

It has been very beneficial to understand how the DLE is updated since it is a more complex process than what is done at the RAE. The RAE has academic sessions and different committees to approve any changes. On top of that, the RAE also interacts with the *Asociación de Academias de la Lengua Española* (ASALE), present all over the Hispanic world.

It is our intention to propose an agreement between academies leading to a systematic labelling of the specialized use of a particular entry/meaning and its representation.

Acknowledgments

I want to thank the ELEXIS project for the opportunity given to me. In the digital age, common standards are highly needed to create structured and interoperable lexical databases that will optimize the lexicographical work. I would like to thank Jordi Porta and Elena Zamora. And also Paz Battaner, who gently came to assist me in my final presentation.

Ivana Filipović Petrović, PhD
Linguistics Research Institute of the Croatian Academy of Sciences

Report
on Elexis Transnational Research Visit Grant at Trier Center for Digital Humanities
(Trier, Germany, February 25 – March 8, 2019)

Travel grant: Call 1
Project title: (Retro)digitisation and online publication of the *Croatian Dictionary of the Literary Language*

Introduction

The project proposal that I submitted to the call for Elexis grants for research visits included the plan of (retro)digitisation and online publication of the *Croatian Dictionary of the Literary Language* (CDLL), as well as computer processing of the corpus on which this dictionary is based. Therefore, the goals of my visit were gain the knowledge, competence and (if possible) lexicographical tools and infrastructure for this project. By the end of my visit, it was clear that not only the goals were accomplished, but so much more: I gained a valuable experience of working with experts in the field of digital humanities and with sophisticated lexicographical tools as well as a social and emotional experience of a two-week living in a foreign country. The hosts, Trier Center for Digital Humanities, especially Dr Vera Hildenbrandt and Li Sheng contributed to all mentioned experiences in a best possible way.

In this report I will present the workflow of my visit, i.e. the projects, dictionaries, tools and all useful insights that had an impact to the achievement of the goals. In addition, I will single out some possible solutions for the further process of (retro)digitization of the *Croatian Dictionary of Literary Language*.

Digitization of primary sources: OCR4all

The first goal was the digitization of the primary sources for the CDLL, in order to build an integrated, fully searchable corpus of 400 sources. For this task, I was introduced in tools and methods of digitization, especially to the open-source tool OCR4all which is based on a deep learning algorithm. Given that we have 400 sources for CDLL stored in .JPEG format, I brought them with me on the external disc. OCR4all works in a way that I upload the source in the .JPEG format and set the digitization of the first few pages in motion. After that, the next step is checking the result and correcting misreading in order to train the program and achieve better results on the next pages of the source. In the case of my first test source, the collection of poems of Croatian writer Tugomir Alaupović, the result on the first three pages was not good enough, due to the poor quality of images as well as the presence of diacritical letters and some special characters in the source. But after the training of a model, the result was significantly better: the accuracy of recognition increased to 98, 7 %. In order to improve the result, it is necessary to train the new model based on the corrections. When it comes to the literary (primary) sources for a dictionary, the result must be 99, 95 %, which can be achieved by OCR4all in combination with proofreading. Furthermore, the result is fully-digitized text, which means that search possibilities are increased (unlike the image digitization that we had), as well as the possibilities for a representation of references in the on-line version of a dictionary. In addition, during my visit I was introduced to the methods and tools for compiling the *Mittelhochdeutsches Wörterbuch* by Dr Niels Bohnert. Among other interesting things (lemma list, dictionary writing system, online version), my attention was drawn by the database of the sources used for this dictionary as well as the list of bibliographic references. Actually, it was a pure example of the database that should be provided for CDLL sources, and that is why it is important to make a full-text digitization of the sources. In conclusion, this form of digitization should be seriously considered when digitizing the corpus of the sources for CDLL. At this point, the OCR lectures and training was completed, and we moved along to the other steps in the process of (retro)digitization.

Encoding in XML according to TEI standards and Dictionary Writing System

The second issue from my project proposal, dictionary writing system refers to the new volumes of CDLL which need to be written. Given that every dictionary has its own requirements and peculiarities, very often the solution is developing an in-house dictionary writing system for particular purpose. Also, some dictionary projects use an XML editing tool and make some tailor-made solutions, i.e. they adopted it for lexicography. The TCDH has experience in the development of a dictionary writing system (TAReS: A Webbased System for Editing, Producing, and Publishing Dictionaries in Distributed Offices) and also solid experience with an XML editor such as Oxygen. I have a very good experience with Lexonomy, given that this open-source platform for writing and publishing dictionaries is currently in use on another project that I am working on (*Online Croatian Dictionary of Idioms*). In order to choose or build suitable DWS it is important to know the structure of a dictionary, but also it would be useful to have some knowledge about the mark-up language such as XML.

Furthermore, encoding in XML will be needed in (retro)digitization process as well, after the digitization of printed volumes. So considering all that, it was very useful for me to learn how to encode in XML for various reasons. I was introduced to the rules of this markup language as well as the TEI guidelines for the encoding of dictionaries. Given that the first 12 volumes of CDLL, that were published between 1985 and 1990, require detailed revision according to the consistent lexicographical treatment established for the new volumes,[1] my practical work on encoding in XML started on the entries from the volume 13 (2013). By the end of my visit, I was able to successfully encode different types of entries from CDLL. I will also be able to use this knowledge if we are going to use GROBID Dictionaries, which is a tool for structuring dictionaries, for converting the data from PDF into the TEI XML, at least on the PDFs of two new volumes that are made from Word.

In conclusion, the benefits of learning how to encode in XML according to TEI are twofold: it will serve as a guide when choosing and building DWS for completing CDLL, and also as the data for an online version of the *Dictionary*.

[1]The *old* volumes of this dictionary suffer from some serious imperfections: the citations often don't match to the source reliably due to the fact that they were rewritten from the index cards into the *Dictionary*, and not directly from the source. Also, the lexicographical treatment of the figurative meaning is not thoroughly implemented. Beside this, bibliographical references in the first 12 volumes are consisted only from the last name of the author, which is insufficient for the contemporary user. All these problems are solved in the new volumes and the *Dictionary* is methodologically significantly improved.

Online publishing

Finally, the third step in (retro)digitization of a legacy dictionary – it's online publication requires the building of a database (with the text and tagging as an output of encoding) and the development of a graphical user interface. In this field, TCDH has some remarkable solutions, for example the interface for the *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm* or the *Goethe-Wörterbuch*. Besides that, in the Center is currently under development an open-source tool for the online publication of dictionaries. This so called Dictionary Viewer will provide an infrastructure that allows publishing a dictionary with its accompanying pretexts like prefaces, abbreviations indexes etc. Also, the *Wörterbuchnetz* is a dictionary network where a user can search for words in 28 German dictionaries (monolingual, bilingual, historical, general and special purpose). All this achievements in the field of online publishing of the dictionaries are very significant and progressive and it was very useful and stimulating for me to be at the heart of them.

Conclusion

At the end of my second week in Trier I started to draw some conclusions about the process of (retro)digitization of legacy dictionaries and what it requires. Moreover, I was thinking about the state of the art on the projects of this kind: projects concerning historical dictionaries, especially the ones that are not finished yet and they started a hundred years ago. The common situation for many of them is the same: you have the material in some form (printed dictionaries, printed sources, maybe scans or images, Word and PDF documents, handwritten index cards). Also, you have a two or three lexicographers with a linguistic background that are busy and overwhelmed with the senses, figurative meanings, examples of use and consistent lexicographical treatment. And the third thing that you have is a desire for technical improvement.

What can you do? This extraordinary group of experts in TCDH taught me and showed me what you can do and I will use these insights for CDLL. When it comes to (retro)digitization of printed volumes, first step is image-digitization of origins. Second step is post-processing of digital images via OCR of this scanned origin and proofreading or double keying of the text. Next step is the analysis of an entry structure in order to start with the markup, i.e. encoding. The last step is using encoded data (XML files) for the building of a database and finally the online publication. In addition, if you have the

sources for the dictionary, which should be digitized as well, you should apply the first two steps and then create a database of the sources.

Some of these steps can be realized by open-source tools. For example, in the case of CDLL we will consider OCR4all, Sketch Engine option *create your own corpus* for our sources and Lexonomy as DWS as well as online publishing, given that I already have an experience of configuring on this platform. Although, it should be pointed out that you still need experts from several fields in order to achieved all this: an expert in computer science and an expert in digital humanities. For two wonderful weeks, I enjoyed the company of some of the best.

Special thanks to: Dr Vera Hildenbrandt, Li Sheng, Dr Niels Bohnert, Dr Hanna Büdenbender, Anja Hennemann and Felix Thielen.

ELEXIS Report

ELEXIS Transnational Research Visit Grant

I decided to apply to this research visit grant in order to observe and learn from lexicographers in loco. Although my research interests are over texts used in specialised-communication context, the major goal of this research visit was to consolidate and share lexicographic and terminological methodologies. Moreover, the opportunity of observing what the 'real-life' of a lexicographer is, and what constraints might come across in a given lexicographic project in the current society of digital information, would assuredly be beneficial to my ongoing project. The Dutch Language Institute (INT) seemed to be the closest working environment to what my project is, thus the reason for choosing this hosting institution. The visit proved to be above my expectations. From historical dictionaries to terminological resources, the in-house projects revealed to be a true inspiration to any lexicographer, corpus / computational linguists, and terminologists, among others. For such a unique opportunity, I am most thankful to the ELEXIS project.

Travel Grant: Call 1

Hosting institution: Instituut voor de Nederlandse Taal

Period of stay: February 4-8th, 2019

Researcher: Margarida Ramos

Affiliation: Centro de Linguística da Universidade NOVA de Lisboa – CLUNL, Portugal.

Current position: PhD Student in Linguistics: Lexicology, Lexicography and Terminology.

Project title: Knowledge Organization and Terminology: application to Cork

ELEXIS Report

ELEXIS Transnational Research Visit Grant

Introduction

I decided to apply to this research visit grant in order to observe and learn from lexicographers *in loco*. The primary goal of my research visit was to consolidate and share lexicographic and terminological methodologies. Moreover, the opportunity of observing what the 'real-life' of a lexicographer is, and what constraints might come across in a given lexicographic project in the current society of digital information, would assuredly be beneficial to my ongoing project.

The twofold structure of this report aims at representing the order of the meetings and topics in a resumed way, while the plain text points at the highlights of the visit, along with some reflexions.

Research goals

My ongoing PhD Thesis project focus is the terminological analysis of specialised corpora resorting to semi-automatic tools for text analysis, in order to systematise lexical-semantic relationships observed in specialised-communication context and subsequent modelling of the underlying conceptual system.

The final goal of the project is to propose a multisemiotic e-dictionary, designed as a multilingual and multimodal product, where several resources, namely linguistic, conceptual, and multimedia are pairing each other to facilitate the user knowledge acquisition. Such an e-dictionary denotes what we consider a useful terminological tool in the current society of digital information.

Writing terminological definitions in natural language is a critical part of my research project, along with the conceptual organisation of the domain under analysis. For that purpose, and given the current lexicographic e-generation, it is expected to work with digital environments which requires from the user certain

Overview

creativity along with its counterpart, the labour-intensive data analysis tasks. It is this merge of creativity with scientific work that motivates my interest on contacting lexicographers and computational linguists, in short connection with informatics, i.e. a multidisciplinary collaborative team.

Finally, TEI XML has proven to be an asset for the lexicographic part of my terminological work. The span of TEI applications goes beyond the perspective of text perpetuation, which leads me to interrogate how far can terminologists go with such encoding text standards, given the underlying reusability and interoperability conveyed by XML environment tools. Thus, the prospect of observing how and what lexicographers use as text processing tools, as well as dictionary writing systems, was another point of interest.

Brainstorming sessions

Meeting lexicographers at their 'real-life' working-context was a genuine opportunity for me, but mostly highly inspirational to my ongoing terminological project.

Most of the lexicographers that I had the opportunity to meet with, at the Instituut voor de Nederlandse Taal (INT), showed how their research outcomes can be shared by different projects within the institute. For instance, data from the new project on Word Combinations, which is specifically targeted at language learners, can potentially be reused in the context of the ANW dictionary (Algemeen Nederlands Woordenboek), a scholarly dictionary of contemporary Dutch, which also includes information on phraseology. Given this scholarly focus, the evidence of how words combine or tend to co-occur, either in a fixed or semi-fixed combination, may thus be considered. This option grounds on the notion that word combination is frequently sought by language learners, rather than by the meaning of the lemma (also known as headword). A student might know how to spell a word but, eventually, will misuse it in discourse context given his/her lack of social-cultural heritage, particularly when it comes to idioms and proverbs. Hence, such an element included in the structure of the article is a useful feature to translators or language learners, for word sense disambiguation.

Day #4

Meeting subject: *Terminology & terminological resources,* with Kinable, dr. Dirk

The Algemeen Nederlands Woordenboek (ANW) & the benefits of technology with Tempelaars, drs. Rob

Day #5

Meeting subject*: Case studies on crowdsourcing for Dutch,* with Tiberius, dr. Carole and Dekker, Peter, MSc

Farewell

(i) *Terminological records in the digital era*

(ii) *From slips to XML, the Historische woordenboeken (WNT)*

(iii) *The Middelnederlands Woordenboek archive* with Kinable, dr. Dirk

Another interesting point observed in the ANW is the various search options made available to users. Queries may be performed starting either from the word or from the meaning of the word, which denotes a terminological vein. This feature is an outcome of the so-called *semagram* project. According to its author, "a semagram is the extensive description of the meaning of a word according to a fixed pattern of semantic categories and properties" (Fons Moerdijk)[1]. Thus, definitions are complemented with additional data, such as properties under the chief-words PARTS, BEHAVIOUR, COLOUR, SOUND, BUILD, SIZE, PLACE, APPEARANCE, FUNCTION and SEX – a 'type template' for the category of animal. In practice, by employing key-words that co-relate with the semantic categories and proprieties of the target word, users may have access to natural language definition(s) and correspondent lemma, despite their (un)knowledge of the latter. Moreover, definitions are pairing with 'hypermedia', namely image, movie and sound: a genuine inspiration for terminologists given the denoted onomasionological and semasiolagical queries possibilities, along with the multimodality of meaning representation.

The ANW is an overwhelming project. Describing it would not be feasible, nor is my intention, which extends to other projects as the overwhelming Historische woordenboeken (WNT), a lexicographic work of 5 generations editors (147 years) – the biggest dictionary in the western world – accomplished into digital format in 2007.

However, the subject of crowdsourcing cannot be left without a word. 'Taalradar' is another in-house project yet crowdsourcing-based and plays a central role within the subjects of neologisms, language variation and blends. Therefore, speakers – the real users of language and creators of new words –, are actively contributing to the general language survey, e.g. the crowd is approached through a survey tool platform, within which their answers are recorded. Once statistically and qualitatively analysed by computational linguists and lexicographers, data will eventually be validated. This project is quite impressive from the perspective of the terminological workflow, in which the identified terms require the experts' validation. There is already an implemented methodology for this subject in our linguistic centre – CLUNL – which leads me pondering how positive a collaborative research venture could be, with the merge of terminological methods and new crowdsourcing technologic trends.

Tools & technology, what else for e-lexicography?

One of the key-words, consensually stated by the INT experts, is 'technology'. Tools and technology are undeniably at the core of the e-lexicographic work, in the current digital era. For textual corpora analysis, Sketch Engine is the elected tool. While as for article editing in the lexicographical and terminological work, in-house tools are developed for the former, and off-the-shelf tools are adjusted for the latter. Article editors are therefore sophisticated lexicographic tools with which experts must

[1] https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/61

deal with daily, but foremost considered an added-value for the effectiveness of their lexicographical tasks and goals.

NedTerm, a terminological resource

Finally, but not less important, I was introduced to the INT terminological resource: 'NedTerm', a platform where users can search for special-field domains and related documentation (e.g. term lists), or hyperlinked towards other terminological resources, corpora, terminological tools, among others, publicly available on the internet. The focus of this platform is consistent with the scholarly profile; thus, valuable tutorial elements can also be accessed by users.

I also had the chance to observe another terminological work, which is being developed by the local terminologist, dr. Dirk Kinable, with whom I had stimulating discussions. Here, the tool used to edit terminological data is QTerm, a well-developed interface, though not open-access, where terms are recorded along with relevant data, such as the elements 'definition', 'POS', corpus evidence, and others. According to dr. Kinable, the final product of this research aims at a descriptive terminology of the domain under analysis, whose concepts are organised through an ontological structure: a tree-format representation of the domain, super- and subdomains. Terms, on their turn, are recorded in the structure of the terminological article, as well as synonyms, if applicable. As discussed, synonyms are common findings throughout the terminological work, thus the necessary *a posteriori* experts' validation of the (preferred) terms.

The tool QTerm is also an interface XML-based. It denotes high potentialities for the terminological work given the wide range of elements possible to create in the article. From here, one can conclude that XML is the ideal environment for e-dictionaries.

Conclusion

Tools for textual data treatment such as data mining, editing, storage, management and publishing, are currently one of the major concerns within the lexicographic work. The use of a common machine-readable language is therefore paramount for the interoperability and reusability of data, given the horde of tools involved in the creation of an electronic language resource. For which XML is unquestionably the predilect format for text digitisation and subsequent web publication.

Still, in truth, it is the multidisciplinary team that makes the dream happen.

Acknowledgements

The visit proved to be above my expectations. From historical dictionaries to terminological resources, the in-house projects revealed to be a true inspiration. For such a unique opportunity, I am most thankful to the ELEXIS project.

I would like to extend my sincere appreciation to all at the INT for the warm hospitality, in particular to dr. Carole Tiberius who kindly assisted me throughout the whole visit, as well as to lic. Lut Colman, for the introductory session: an interesting project over word combination; to dr. Tanneke Schoonheim, from whom I learn historical facts of Leiden City and of the beautiful building where the INT is located; to lic. Katrien Depuydt, for the relevance of corpora metadata; to dr. Rob Tempelaars, for the explanation of the lexicographic article edition and the ingenious 'semagram', in the ANW; to MSc Peter Dekker for his inspiring crowdsourcing project; and to dr. Dirk Kinable, who, beyond all stimulating terminological discussions, also showed me the precious Middelnederlands Woordenboek archive, as a Farwell-gift after my enthusiasm on old dictionaries. Finally, to prof. dr. Frieda Steurs, Head of the INT, who kindly accepted my application at theirs and enthusiastically motivated me to make a presentation of my terminological project to the team.

Copenhagen Business School
HANDELSHØJSKOLEN

Department of
Management, Society and
Communication

Dalgas Have 15
DK-2000 Frederiksberg

Tel: +45 · 3815 3815
Fax: +45 · 3815 3830
www.cbs.dk

15 December 2018

Henrik Køhler Simonsen
Tel.: +45 9399 1489
E-mail: hks.msc@cbs.dk

Report on Transnational Research Visit Grant at K Dictionaries - 6 to 13 December, 2018

It is with great pleasure that I submit this report on my Transnational Research Grant Visit at K Dictionaries from 6 December to 13 December 2018.

My research visit at K Dictionaries was extremely useful, educational and inspirational and has accellerated my research in business modelling and strategies in lexicography.

My cooperation with the host and the host organisation was very good in all phases of the project and it has been a great experience for me.

I am therefore pleased to briefly describe the project and the results that we achieved during my stay at K Dictionaries.

Should you need further information, I would be pleased to provide further information if needed.

Yours sincerely,

Henrik Køhler Simonsen
External Lecturer, PhD, MA, MBA

Department of Management, Society and Communication
Copenhagen Business School
15 Dalgas Have
2000 Frederiksberg
Denmark

Report on Transnational Research Visit Grant

Introduction:

The research objectives of the project *Business Models & Strategies for Lexicography* were to explore new methods, new technologies and new partnerships, which may eventually become proper business models. I chose K Dictionaries for my project, because K Dictionaries was the only privately owned company of the eleven hosting institutions and infrastructures, and I know from research already conducted that privately owned companies in lexicography are facing significant strategic and financial challenges.

Research objectives:

The research objectives of the project *Business Models & Strategies for Lexicography* project include examining, recording and generating new insights, ideas and strategies on alternative business models for lexicography, including reflections on new technologies, platforms, user groups, applications, eco systems, revenue streams, and interoperability with other domains. The more concrete research objectives were to conduct a large number of research interviews and to arrange and facilitate a number of strategy workshops and to video record these workshops for subsequent analysis.

Research Project Phases:

The research project involved the following phases:

Before:

- Preparatory meetings with host organisation
- Preparatory interviews with relevant experts, businesses and companies
- Practical arrangements (interviews and planning of strategy workshops)
- Preparation of interview guides
- Scheduling of research interviews
- Conduction of research interviews

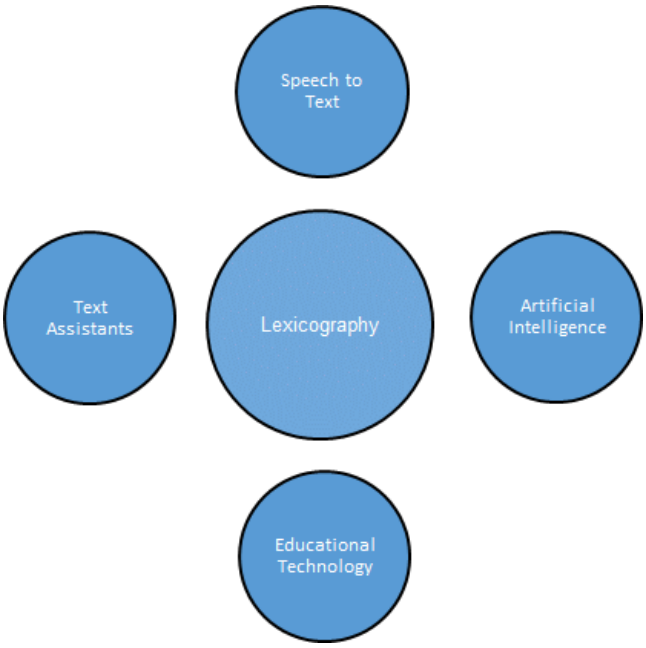
During:

- Internal strategy workshop on Strategy & Value Propositions at and with K Dictionaries
- Internal strategy workshop on Sales & Marketing at and with K Dictionaries
- Internal strategy workshop on Business Model elements at and with K Dictionaries
- Business model innovation workshop with speakers and participants from Israel, Denmark, Norway and Portugal
- Scheduling of research interviews
- Conduction of research interviews

After:

- Conduction of research interviews
- Analysis of interview and seminar data
- Writing report to ELEXIS on the project
- Writing an article to Kernerman Dictionary News

Overview of interviews:



A total of 30 interviews with leading experts from the five areas shown above were conducted during the three project phases.

The interview data and the seminar data have not been analyzed nor categorized yet, but selected insights from the interviews will probably be published in an article in Kernerman Dictionary News.

Preliminary results:

It is important to remember that it is not possible to develop a one-size-fits-all business model. This is a very complex question and each market, service, country, language and technology should be taken into account. However, my research seems to have produced the following insights that may eventually be developed into proper business models. The many interviews and the video recorded workshops seem to indicate that lexicography in privately owned businesses could be business developed by focusing more on:

- Ubiquitous, integrated, automated, effortless and flow facilitating lexicographic services, such as for example text writing assistants Grammarly, Textio or Write Assistant
- Mobile and domain-specific lexicographic services, such as for example the mobile services provided by Visioneducation or Clarify
- Community or Cloud Funded lexicographic services, such as for example Wordnik
- Domain-specific and corporate-focussed lexicographic services, such as for example Altomhus.dk
- API-delivered lexical data sets, such as for example Lexicala API
- Learning-integrated lexicographic services, such as for example MV-Nordic, Encyclopedia Britanica etc.
- Lexicographic services based on new technology such as Artificial Intelligence, Speech toText and Language Model, such as for example Dictus, Skype Translator etc.

6 Concluding remarks on the outcome of year 1

Overall, year 1 has been highly successful as far as the execution of the first round of ELEXIS visiting grants is concerned. All five grant holders reported on high scientific and technical value of their visits. As evidenced by the reports, the host institutions have provided the necessary guidance and support, both scientifically and technically, as to help the grant holders move forward in their projects. The topics span from detailed insight into micro- and macro-structures of contemporary dictionaries and procedures for retrodigitisation, to business models for dictionaries, terminology, and visualisation of language variation.

At a later stage when more grant visits have been completed, it would be interesting to make a review of outcomes also from the perspective of the lexicographical milieus that have hosted the researchers. Hopefully, the knowledge gain goes both ways. One issue has arisen, and will only increase during ELEXIS' lifetime: namely, some host institutions are very popular and receive many applications for visits, whereas other institutions have not yet received any application. It should be considered whether initiatives should be taken to calibrate the budget better to avoid this situation. Alternatively, strategies for better promotion of host institutions that attract less interest need to be considered