elexis european lexicographic infrastructure

5.1 ELEXIS SKILLSET REPORT

Author(s): Toma Tasovac, Monica Monachini, Fahad Khan

Date: 31.01.2019

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.

elexis european lexicographic infrastructure

H2020-INFRAIA-2016-2017
Grant Agreement No. 731015
ELEXIS - European Lexicographic Infrastructure

5.1 ELEXIS SKILLSET REPORT

Deliverable Number:	5.1
Dissemination Level:	Public
Delivery Date:	31.01.2019
Version:	1.0
Author(s):	Toma Tasovac Monica Monachini Fahad Khan

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
0.5 (15.12.2018)	Intro, Aims & Methodology	Toma Tasovac
0.9 (29.01.2019)	Analytic Summaries	All three authors
1.0 (31.01.2019)	Executive Summary	Toma Tasovac

Executive Summary

The ELEXIS Skillset Report is based on in-dept interview with 11 lexicographers across Europe, which were conducted in order to provide a) a general assessment of the skills required for the active participation in the ELEXIS research infrastructure; and b) a set of recommendations about the topics and modalities of the training materials to be developed by ELEXIS.

The following summary presents key findings and recommendations that were identified in a qualitative assessment of the interviews:

- Almost all of our informants received no formal lexicographic training.
- Almost all of our informants describe themselves as lexicographic self-learners who honed their lexicographic skills by:
    - consulting various dictionaries
    - reading scholarly literature
    - attending workshops and seminars outside the regular university curricula; and/or
    - learning from their colleagues or on the job.
- Investing in continued skills development lexicographers, regardless of their level of experience, is considered to be of crucial importance.
- Developing new lexicographic skills faces various challenges including:
    - Lack of university curricula in lexicography or systemic, integrated training outside the university
        - *The knowledge gained in unconnected workshops is partial and segmented* (Hudeček)
    - If lexicography is taught as a university subject, it is in some countries too theoretical and removed from contemporary trends and best practices
    - If students have a background in linguistics, they don't necessarily have the necessary technical skills, and vice versa
    - Technology is not the ultimate challenge: reluctance to adopt new methods is also a social and cultural challenge
    - Not all lexicographers believe in a culture of open access, open source and sharing lexicographic data.
- Specific topics that should be covered by ELEXIS training materials ought to include:
    - fundamental theoretical concepts in lexicography and terminology, including:
        - lexicographic macrostructure
        - lexicographic microstructure
        - dictionary typologies
        - onomasiological vs. semasiological models
        - lexicographic processes
    - fundamentals of data modeling and text processing, including
        - structured data and markup languages (XML, TEI, MDF)
        - abstracting information structures across a large number of dictionary entries

        - differentiating between representation (of an abstract model) and presentation (or a typographic realization) of the said model on paper or computer screen
        - scripting languages for text processing
        - community and official standards such as TEI and LMF
    - fundamentals of corpus linguistics
    - fundamentals of lexicographic practice
        - including common though difficult tasks such as:
            - selection of entries
            - splitting senses
            - formulating good definitions
            - choosing adequate examples
        - tools such as
            - dictionary-writing software
            - XML editors such as oXygen
            - POS taggers
            - parsers
            - automatic term and collocation extractors
        - and user-centered approaches to
            - dictionary usage research
            - using dictionaries in teaching and learning
            - user participation and crowdsourcing
    - fundamentals of usability and design
    - fundamentals of intellectual property rights and copyright issues
- Teaching tools is important, but teaching critical use of tools even more so
    - *The tools of today are not necessarily the tools of tomorrow. We should prepare students not only how to learn tools but also to be able to switch to different tools.* (Costa)
- Face-to-face is the preferred teaching and learning method
    - presential settings facilitate communication between instructors and students
    - intensive face-to-face workshops provide a dedicated space for learning, without the usual distractions of day-to-day obligations
        - *Face to face training is the best. In the five days I spent in Berlin, I managed to digitize a dictionary from scratch without any previous knowledge of GROBID dictionaries.* (Mickoski)
    - learning is a social activity
    - infrastructural services can benefit from being tested in face-to-face settings
        - *We see a lot of different real-life examples [when we conduct training measures] so we can make sure that what we put in TEI Lex-0 reflects the expectations and constraints of the projects we are dealing with in the classes we teach.* (Romary)
    - Face-to-face training is expensive and time-consuming. One model that can be used to overcome this obstacle is *blended learning*, an approach which combines online training materials with traditional face-to-face opportunities.

- ELEXIS should consider specifically targeting participants from lesser-resourced countries as candidates for face-to-face training in order to help overcome the digital gap
- Most informants see ELEXIS as a kind of central hub or reference point for collecting and developing training materials: this should be in the form of a "specific website" or a platform (such as Moodle). The training materials should be on a wide range of topics and openly accessible to everybody.
    - *ELEXIS should contribute towards setting a general training standard as well as a reference point for lexicographic practice in general.* (Grønvik)
    - *ELEXIS could play an important role in integrated curriculum development. (Hudeček)*
- Training materials should
    - be explicit about their learning objectives and outcomes
    - be explicit about the target audience
        - beginners
        - advanced learners
        - experts
    - maintain a balance between theory and practice: both are important and necessary, but theory should not trump practice
        - *Lexicography is very practical, and it would not be that useful to have an abundance of theoretical training material* (Klosa-Kückelhaus)
        - *No matter how much lexicographic theory you learn, it doesn't help unless you actually start producing dictionaries* (Mickoski)
    - convey the fun aspects of working on dictionaries without masking the difficulties involved.
- Many informants prefer multimodal training materials, although some still prefer written step-by-step instructions
- Videos in training materials:
    - should be short rather than long;
    - should be carefully prepared to satisfy concrete learning objectives;
    - should be accompanied by full transcripts and/or subtitles;
    - should not be served on their own but embedded and textually contextualized.
- Catering to users coming from different backgrounds will be a significant challenge for ELEXIS in terms of training and education. In order not to leave anybody behind, the ELEXIS training materials should not assume any previous knowledge and should cover the very basics, also for non-experts:
    - *Do not assume that anybody knows anything. Think how low you can go.* (Hudeček).
    - *Teachers sometimes forget that they had studied for twenty years and that their students are novices.* (Costa)
- ELEXIS should consider creating a glossary of e-lexicography terms which would help users starting to work using digital tools in their day-to-day work
- ELEXIS should consider building lasting communities around training measures and training materials
    - One possibility to do it would be through an online-forum, in which responsive members of the community can help each other, while the authors can get feedback on their training materials

        - *Whenever you come across a problem, it's good that you can find some other people who already had encountered the same problem.* (Özcan)
    - Another possibility would be to create a create a blog with exercises to be completed after the training
        - *It's great to have a training week, but after the training is over, the students are alone again and without support.* (Salgado)
- ELEXIS should consider experimenting and adopting creative approaches to teaching which could include:
    - Lexicographers "thinking aloud" and describing what they're doing as they carry out various tasks
    - Video reports of visits to the "dictionary factory"
    - Matching users from lesser-resourced communities with their more advanced counterparts
    - Creating blog posts to accompany and follow training measures
    - Sharing annotated dictionary samples
    - Sharing examples of successful case studies including analysis of various dictionaries

Introduction

Infrastructures need users, and those users who are well-trained in the services provided by the infrastructure will be in the position to benefit the most from the infrastructure itself.

As an infrastructural project, ELEXIS is both a knowledge network and a service provider. It caters to a diverse userbase of lexicographers, with varying competences, interests, and access to technology, across many national borders. With such a diverse target audience in mind, it will be especially important for ELEXIS to invest time and effort into training and educating its potential users as a way of securing the social sustainability of the emerging research infrastructure.

In the first year of the project, WP5 already delivered some initial training measures. ELEXIS, for instance, sponsored the "Terminology and Lexicography" strand at the Lisbon Summer School 2018, which was held from July 2nd – 6th 2018 at the Nova University of Lisbon. During the Summer School, three courses were offered:

1. *Semantics in Terminology: The Contribution of Concept to the Meaning of Term* (taught by Christophe Roche from the Université Savoie Mont-Blanc, France)
2. *Trends and Advanced Approaches in Lexicography* (taught by Teresa Lino, Raquel Amaro, and Rute Costa from the NOVA CLUNL, NOVA FCSH, Portugal); and
3. *From Print to Screen: The Theory and Practice of Digitizing Dictionaries* (taught by Toma Tasovac from the Belgrade Center for Digital Humanities, BCDH).

ELEXIS also partnered up with the Digital Research Infrastructure for the Arts and Humanities (DARIAH) to deliver a Lexical Data Masterclass, which took place at the Berlin-Brandenburg Academy of Sciences from 3rd – 7th December 2018. The masterclass was run by Toma Tasovac (BCDH) and Laurent Romary (Inria, France) and offered instruction in TEI, XPath, XSLT as well as GROBID Dictionaries. The participants came with their own datasets and applied the newly learned skills on their own materials.

In parallel, members of the WP5 worked on evaluating the skills researchers need in order to benefit from and contribute to the emerging lexicographic infrastructure. This report is a product of that research.

Aims of this report

The aims of this report are two-fold:

1. to provide a general assessment of the skills required for the active participation in the ELEXIS research infrastructure and use of its offers and services with particular attention to the issues of the digital gap between the lesser and better-resourced countries; and
2. to provide a set of recommendations about the topics and modalities of the training materials to be developed by ELEXIS.

This report should be seen as complementary to ELEXIS D1.1 *Lexicographic Practices in Europe: A Survey of User Needs.* Unlike D1.1, this report has a narrower focus (skills development and training and education) and is based on a qualitative (rather than quantitative) methodology.

Methodology, Structure and Perspectives

The main basis of this report are in-depth video interviews which were conducted with 11 lexicographers from across Europe, both from within and beyond the ELEXIS Consortium.

Our informants were selected based on two criteria:

1. their professional expertise in lexicography and related fields; and
2. their familiarity with the training and educational landscape in their own communities.

While other methodological approaches (including surveys and statistical data) would without a doubt provide additional insights into the topic of this report, a qualitative assessment of the in-depth interviews has given us a unique chance to highlight in greater detail the complex needs and expectations of European lexicographers when it comes to lexicographic training and education, in general, and ELEXIS, in particular.

Even though this report does not pretend to be in any sense representative, our informants were chosen in a way to assure a number of different perspectives (digital-born vs. retrodigitized dictionaries, lexicography vs. NLP, university professors vs. PhD students, lesser resourced vs. more advanced scholarly and linguistic communities).

The interviews were conducted remotely, using Zoom and Skype, in December 2018 and January 2019. Each interview lasted on the average around 60 minutes. The interviews were recorded. The interviewers followed a concrete interview script (see Appendix I), which contained the most important topics for discussion, while allowing each interviewer to pursue more detailed exploration of the topics, depending on the answers given by the informant.

Each interviewee signed a consent form (see Appendix II).

The recorded interviews were then transcribed and analyzed as a basis for compiling this report.

Each interview section in this report consists of five sub-sections:

1. information about the interviewee's professional background
2. information about the interviewee's educational background
3. summary of the interviewee's main points regarding skills development
4. summary of the interviewee's main points regarding teaching methods
5. summary of the interviewee's main points regarding training materials

The Executive Summary above includes the most relevant and interesting findings of this report. No effort was made to reconcile different views or prioritize among the suggestions offered. This is the work that should be taken up by ELEXIS as it continues to develop, expand and improve its training and education activities.

Interviews

Klara Ceberio Berger

Professional Background

Klara Ceberio Berger is the head of Linguistic Resources at the Language & Technology Unit of Elhuyar Fundazioa, a foundation dedicated to science, technology and the development of the Basque language. The main foundation's commitment is towards quality at all times in order to turn Basque into a language for communication in society. The Language and Technology Unit's main aim is to create and offer language resources, tools and services for the Basque language, and to provide a high-quality, integral language service on the basis of customers' language needs.

Ceberio Berger works on lexical databases for spelling and grammar checks and other types of applications used in the area of natural language processing (NLP). Her research interests mainly focus on lexicography. In particular, she works on updating lexicographical and terminological databases. She has been also involved in different NLP projects, such as corpus building and tagging.

Educational Background

Ceberio Berger studied German Philology at the University of Salamanca in Spain. The degree was more focused on German language and literature. She received no formal lexicographic training, basing her lexicographic experience mostly on consulting various dictionaries.

She made her first steps in lexicography while enrolled in a postgraduate course in computational linguistics at the University of the Basque Country, which was closely related to the IXA Research Group. She first started working as a PhD student for the IXA Research Group where she developed her lexicographic skills on the job. Later, she continued to learn when she started working at Elhuyar. Learning on the job was challenging for her.

Skills Development

Ceberio Berger has been involved with both digital-born and retro-digitized dictionaries. The Elhuyar database was mainly intended for publishing printed dictionaries, but later on some of the features of the database were modified and updated to optimize it for the born-digital content.

According to Ceberio Berger, most people still develop their lexicographic skills "at work". Elhuyar is not a teaching institution, but the foundation invests in continual skills development of their staff: as soon as a new colleague is employed, those responsible at his or her unit will conduct on-the-job training.

Ceberio Berger complained that it was not always easy for her and her colleagues to find the adequate training in the area. They follow and attend various workshops in Europe devoted to lexicographic training. One of the main obstacles to the development of lexicographic skills remains the fact that "it is sometimes not easy to find the information needed for your specific work."

As a result, it happens that people with no teaching experience sometimes need to teach their colleagues.

In this respect, a research infrastructure project like ELEXIS can play an important role in helping lexicographers develop new skills and overcome the obstacles mentioned: ELEXIS could offer courses about lexicography, both the theory and practice in this area and promote inter-institutional visits to know how this work is conducted in other institutions and countries.

Teaching Formats

Ceberio Berger has clear ideas about teaching formats for developing new skills: she recommends three types of courses: continuing education courses that teach the state-of-the-art of lexicography; online courses about different hot topics; and face-to-face intensive courses.

Training Materials

Ceberio Berger points out that there is a need for high-quality training materials that are accessible at various levels. Having a specific website in the form of an open platform that gives access to a wide range of articles and material of interest in this area would be desirable. In particular, she suggests the use of platforms such as Moodle, which can be reinforced by workshops or face-to-face courses. She promises to be one of the first users of the training material created on this topic.

Rute Costa

Professional Background

Rute Costa is a full professor at the NOVA University of Lisbon and President of the Linguistics Research Centre of the Universidade NOVA de Lisboa (CLUNL). At the undergraduate level, she teaches terminology for translators and linguists. She also teaches in the Masters's Program in Terminology focusing on theories of terminology and terminology and ontologies. In addition, she teaches in two PhD programs: one in linguistics, where she teaches lexicology and lexicography, and one in translation and terminology, where she mainly teaches terminology and technical communication.

After she was elected president of the European Association for Terminology, she became an EAT liaison to ISO TC37 in 2001. In the beginning, she worked on more theoretical standards such as ISO 704 and ISO 1087. Now she is president of the TC37 SC2, which is a committee working on terminology workflows and language coding, but she is also involved with SC4 (Language resource management).

Educational Background

Costa studied Portuguese and German Languages, Literatures and Cultures. In addition, she obtained a four-year Teacher's Degree focusing on the methods of teaching Portuguese and German Literatures. She obtained a Master's Degree in Linguistics, Lexicology and Lexicography. She focused on word formation in the economic domain.

After this, she spent some time at the University of Laval and the University of Montreal in Canada, as part of Francois Daoust's program ATO (Analyse de texte par ordinateur), where she honed her NLP

skills. She also worked with Zampolli and Calzolari in Italy. In 1991 or 1992, she brought Lou Burnard, Michael Sperberg-McQueen and Nancy Ides to Portugal to teach TEI.

In her PhD, Costa built a tagger for Portuguese, worked with remote sensing and she built an extractor for multiword terms. She's to this day interested in the differences between collocations and multiword terms. This is still an important topic in terminology but also in lexicography.

Skills Development

Budding lexicographers should have a **solid foundation in linguistics** (especially morphology, lexicology and corpus linguistics). Speaking several languages is not necessarily enough:

> *There is a difference between language professionals such as translators, for instance, who know languages well, and linguists who's job is to think about language.*

Equally important is **familiarity with existing dictionaries** and different lexicographic outputs. How should the data be organized for general users, for students, or in dictionaries for special purposes.

Tools are important and knowing how to use various tools is an indispensable skill, but **we should teach students also to be critical in the use of tools**:

> Tools are not magic. When I teach tools, I normally don't care about the tool, whether it's AntConc or SketchEngine. What I want people to know is how to think about language, how to look at and analyze the data, and how to organize that data in dictionaries. **The tools of today are not necessarily the tools of tomorrow.** We should prepare students not only how to learn tools but also to be able to switch to different tools in their careers.

Finally, it's very important for students that they become familiarized with existing **standards**.

The biggest challenge in helping students develop new skills comes from the fact that our students have different backgrounds. According to Costa, students (or colleagues) with a traditional humanities background are often resistant to learning about technology and standards.

> *In my experience, people from the linguistics community do not like the word standards and the notion of formal thinking. When they hear the word standard, they think we're not letting people use language the way they want to. So we have to explain that's not what we do.*

Also, some habits are hard to break:

> *I still have colleagues who work on dictionaries using Word. That still happens a lot.*

If you try to encourage them to learn new tools, they often say "Oh, but I'm used to working this way." That's why the focus should be on introducing tools and standards to students at a younger age.

Teaching Formats

Costa stressed the importance of learning in international contexts, attending workshops in different countries, expanding one's scholarly and cultural horizons but also having a measure of one's knowledge. International workshops are important not only because one learns particular skills, but

because -- in interaction with other participants -- one also becomes aware of what one still needs to learn.

Costa prefers face to face training, but is also a fan of blended learning, an approach which combines online training materials with traditional face-to-face opportunities.

> *You could do a course in which 80% of the work is done through e-learning, but the remaining 20% requires physical presence.*

Blended learning methods can be very efficient because students come prepared and really take the advantage of the interaction with the instructor in the classroom, asking questions, clarifying issues etc.

Training Materials

When preparing training materials, we should be very **explicit about the skills** that we want to teach and the **learning objectives and outcomes** we want to see.

Ideal training materials are multimodal: they contain written instructions but also make use of images, videos etc. Videos, however, should be prepared carefully and specifically to satisfy concrete objectives. Spontaneous and too general video recordings should be avoided.

In addition to having clear objectives, ELEXIS training materials should include a **theoretical introduction**, which shouldn't be too long, but which should be broad enough to explain to the students that there are different ways of doing things and that there are different theoretical backgrounds to everything.

Finally, the ELEXIS training materials should teach **concrete, practical skills** including handling various tools. They should also include **practical exercises**.

According to Costa, it is very important for instructors -- including those teaching in the training measures that are organized within ELEXIS or preparing training materials -- not to have any presuppositions regarding their students' previous knowledge.

> *Teachers sometimes forget that they had studied for twenty years and that their students are novices.*

Oddrun Grønvik

Professional Background

Oddrun Grønvik is a former editor and then chief editor (2006 – 2016) of the *Norsk Ordbok*. The Norwegian Dictionary is the dictionary of the Nynorsk writing language and the Norwegian dialects, prepared at the University of Oslo (1939–2016). The dictionary was published in print form by The Norwegian Sami Society, financed by the Ministry of Culture and the University of Oslo. The dictionary consists of twelve volumes and is the first scholarly dictionary documenting the Norwegian language. Grønvik also worked as an Associate Professor at the University of Oslo, and is now retired. She was

also a consultant for the Norwegian Language Council (1979-1987), and was engaged by the University of Bergen in the start-up phase of lexicographical projects at UiB (Revision *Bokmålsordboka* and *Nynorskordboka*; Revision *Norsk Ordbok a-h*).

Grønvik's main tasks, apart from editing the assigned alphabet sections, has always had to do with the overall organization of scholarly lexicographical projects, from material collections, source systems to quality control of the finished manuscript, as well as training the personnel (lexicographers and research assistants).

She was also involved in lexicographic projects with African languages, especially in Zimbabwe (The Allex Project 1991 – 2006 and the CROBOL Project 2007 - 2011), with a particular focus on the organizational side, i.e. helping to administrate lexicographical projects, establish courses etc.

After 2002, *Norsk Ordbok* was reorganized on a digital platform, and a much tighter overall organization was set up. Grønvik's main activities in this phase included interacting with the ICT unit responsible for writing the software and providing theoretical training to new lexicographical staff. She was the instructor of a training course in scholarly lexicography, which consisted of 12 x 2 hours and was aimed at people with basic research training. She was also responsible for writing the editorial guidelines. Currently, she is involved with both lexicographic theory and practice, through the projects she is a part of at the University of Bergen, and partly through various researcher-defined smaller projects.

Training Background

Grønvik has a B.A. Hon. in English Literature and Language from St. Hilda's College, Oxford (1970) and a Cand. Philol. from the University of Oslo (1985). In 2006, she was conferred a Dr. Litt. degree *honoris causa* from the University of Zimbabwe.

As far as lexicography is concerned, Grønvik says that her generation had to learn Latin to start language studies at the University in Oxford and in Norway. She studied three foreign languages (English, German and French) in her secondary school, plus Old Norse as part of her mother-tongue studies. The use of dictionaries and reference works was taught at the time, but lexicography as a university subject was not available.

The University of Oslo was in 1971 tasked by Parliament to house the national language and name collections as well as the reference works based on them. A special institute was created for that purpose – the Norwegian Lexicographical Institute. *Norsk Ordbok* was housed there. *Norsk Ordbok* integrates spoken and written source materials and gives a broad coverage of dialect variations, both diachronic and synchronic. Grønvik had the opportunity to learn the tradition of highly organized philology through editing dictionary manuscripts available in the the Norwegian Language Collections. Her teachers were older colleagues. Once employed at the *Norsk Ordbok,* Grønvik looked for courses, and attended her first lexicography workshop in 1988, arranged by the *Ordbog over Det Danske Sprog*. Then, she attended the first TEI conferences and professor Hartmann's course in Essex (1988). She also edited a book on Norwegian lexicography together with a colleague.

Interestingly, Grønvik pointed out that ***lexicography*** in Norway – and in all the Nordic countries – ***has a vital connection to language standardization*** and to some extent to developing literacy, especially in

the minority languages of the Nordic countries. Another important aspect of the language council work is ***bilingual lexicography between the Nordic languages,*** and also to some extent between each Nordic language and the recently arrived minority languages. This is a prime example of **lexicography as a tool in social integration**.

Grønvik's most memorable learning experience was "without doubt the first phase of the ALLEX Project (1991-1996), which provided the Shona with their first monolingual mother-tongue dictionary." The project stemmed from a cooperation between Norwegian, Swedish and Zimbabwean lexicographers and computational linguists. From the Nordic end, the ALLEX Project served to prove that lexicography could be taught as an academic discipline, and that linguists could be trained to become lexicographers from scratch and at speed. An adequate first monolingual dictionary for an African language could be produced within a five-year time-span, given the resources and the organization needed.

According to Grønvik, "the operationalization of linguistics into lexicography, in a fashion both efficient and scholarly" was her defining professional experience.

Skills Development

In terms of the skills needed to create digital-born dictionaries or retrodigitize legacy dictionaries, Grønvik highlighted the following:

1. In terms of digital-born dictionaries, one has to have a **proper and deep understanding of the category systems deployed in linguistics**, and on that basis **develop a model for a) the dictionary and b) the entry**. And once the model is decided on, one has to stick to it until the project is finished.
2. In terms of retrodigitizing dictionaries, **one has to uncover the model that structures the text**. One also has to treat the original work as a text, and not try to force it into a more stringent format (for instance, a totally consistent relational database format). **Philology has to come first.** Indexing older lexicographic works must be done with a proper understanding of the fact that the indexing added represents a **present-day interpretation of the older text**, and **it must always be possible to see the older text in isolation**.

To develop those lexicographic skills, **theoretical teaching and reflection**, combined with **trial-and-error experiences working on real projects** is exactly what's needed.

All the projects that she has worked on have offered in-house training, both in terms of theory and monitored hands-on editing of assigned alphabet sections. There has also always been a formal quality-control system by which manuscripts are read, commented on and decisions added to the editorial guidelines as a running process. For her, **theory courses and monitoring support through the first year of working as an editor** is the best way to help lexicographers.

Investing in continued skills development of the staff is of crucial importance. The following ingredients are essential: conference participation every year, especially lexicography conferences; writing articles on request from the chief editor or on one's own initiative; in-house seminars: the *Norsk Ordbok* organized a two-day seminar with invited and in-house speakers twice a year; participation in seminars and conferences organized by neighboring professional groups (within corpus linguistics, semantics, grammar, dialectology etc).

Grønvik considers the last item particularly important in integrating lexicography projects in their surrounding linguistic environment and preventing professional isolation.

The teaching and practice of lexicography, according to Grønvik, operate best in supportive surroundings and with a measure of established standards of quality and performance.

Because lexicography is extremely labor-intensive, if projects are to operate efficiently, the following should be ensured:

a) easy and systematic access to the chosen raw materials (corpora, paper and digital language collections;

b) a group of people working together on the same or similar projects (self-training is for the few); and

c) a project framework giving direction to the work (deadlines, some outside pressure expressing the need for the planned product).

Grønvik therefore thinks that *long-term institutionalization is essential, both in relation to teaching and training, development of language collections and pushing forward large-scale or smaller lexicographical projects*. The most destructive thing a language community can do, is to individualize responsibility for carrying out (sections of) lexicographical projects.

In this context, ELEXIS has the potential to help create and expand professional networks. ELEXIS should contribute towards setting a general training standard as well as a reference point for lexicographic practice in general without making unrealistic demands on uniformity: "A general training standard is no hindrance to local or project adaptation."

Teaching Formats and Training Materials

Grønvik would like to see the development of a web-based training course, a sort of "open university introduction to theoretical lexicography and lexicology" which could be supplemented locally with materials for lexicographic practice. There could be a basic course corresponding to a standard one-term course at the BA level (e.g. 12 double lessons with questions to answer and papers to deliver), followed by various add-ons, for instance, corpus building, terminology, dialect lexicography, field work collecting oral materials etc.

All of these topics have a general basis which would be applicable to any language or language pair.

Some web-based courses are restricted in a way that you have to complete one section before you start the next. This is something that should be be considered.

There should be some kind of certificate to be obtained. To avoid setting up new institutions, the certificate could perhaps be issued by the institution where the student is enrolled.

Antton Gurrutxaga

Professional Background

Antton Gurrutxaga is a lexicographer and NLP researcher at the Language & Technology Department of the Elhuyar Fundazioa, a non-profit Basque organization that was initially set up as a cultural association in 1972 before becoming a foundation in 2002. The Elhuyar Fundazioa is largely devoted

to promoting the development of the Basque language and of popularizing science and technology. The organization is also involved in the publication of dictionaries, both of the general and the specialist varieties.

Gurrutxaga specializes in the development of computational tools and resources for the Basque language. He works on tasks such as the creation and annotation of corpora as well as on dictionary building. With respect to the compilation of dictionaries, he has been involved in the compilation of specialist Basque dictionaries of scientific terminology as well as a Basque-Spanish bilingual dictionary, the latter of which is the most well-known of all the dictionaries produced by the Elhuyar Fundazioa in the Basque Country, and which has proven to be especially popular in its online version. Gurrutxaga is also involved in a collaboration with the Euskaltzaindia, the Royal Academy of the Basque Language, and works on the creation of corpora and corpus-analysis tools that support the compilation of Euskaltzaindia's monolingual dictionary of Basque.

Educational Background

Gurrutxaga received a PhD in Computational Phraseology in 2014. His dissertation was entitled "Automatic characterization of the idiomaticity of noun-verbs expressions in Basque." One part of his doctoral work was a theoretical study into the kinds of phraseology that exist in the Basque language; another part was more practical. Interestingly, Gurrutxaga studied chemistry at the university, but then changed over to Basque Language Studies. He feels that his early scientific training helps him in his work dealing with scientific terminology.

The possibility of studying lexicography or terminology at a university level was and is still not available in the Basque Country. So that even though he received training in designing lexical databases during his Master's in Language Technology, this was mostly oriented towards computational uses, rather than with an eye to compiling dictionaries. In his case, much of his training as a lexicographer was on the job and was strongly informed by the practical necessities of publishing dictionaries. Indeed, due to the fact there wasn't enough staff to provide in-house training at the Elhuyar Fundazioa when he initially started working there, he largely had to organize his own learning, i.e., by reading articles, following seminars, and attending a number of conferences or masterclasses. One memorable training experience for him was the Lexicom Masterclass held in Brno and taught by Sue Atkins, Michael Rundell and Adam Kilgariff.

Skills Development

According to Gurrutxaga, the essential skills for lexicographers include 1) a firm grasp of the fundamental theoretical concepts in lexicography/terminology along with a good knowledge of the language(s) involved in the resource(s) in question; 2) a basic understanding of corpus building and above all the use of corpora in the compilation of dictionaries. In the case of the Basque language, textual corpora have only recently begun to appear and these have allowed lexicographers to refer to cases of authentic language use in the compilation of dictionaries; at the same time, tools such as POS taggers, parsers or automatic term and collocations extractors have also been developed for Basque, and are currently used in corpus analysis and dictionary making. This is quite different from the situation when Gurrutxaga started working on creating terminological works for Basque in areas such as chemistry or mathematics, which at the time had to be derived from scratch.

Gurrutxaga also considers it important to train people in such common though difficult tasks as entry selection, deciding when to split senses as well as the formulation of good definitions and the choice of adequate examples.

Gurrutxaga mentiond a two-year-long Master's course in Terminology offered by the University of Pompeu Fabra, which two of his colleagues attended and which they found useful. The problem with university training however is that it can be costly, and his institution currently does not have enough funds for it to be a long-term solution in respect to fulfilling their training needs.

Teaching Formats

Gurrutxaga recommends the use of materials in format such as written articles, video interviews with lexicographers, bibliographic materials, and video tutorials that would be accessible from a platform such as Moodle. The provision of an online platform which would allow users the freedom to access lexicographic materials whenever was most convenient for them.

Training Materials

Gurrutxaga pointed out that there is a need for high-quality training materials that are accessible to non-experts: this is especially the case given the recent increase in people who are interested in publishing online on sites such as Wiktionary, but who, at the same time, lack the fundamentals of what a dictionary is, what a part of speech is, or how to write a definition.

Having an open platform that gives access to tools, tutorials on how to use them and information on successful case studies would be especially desirable for lesser-resourced languages such as Basque and would allow more people to get involved in contributing towards the production of lexical resources. Gurrutxaga believes in the importance of maintaining a balance between more practical training materials, those that focus on carrying out specific tasks, and those that are more theoretical. Having a good theoretical background allows lexicographers to be more independent, but the practical side is crucial too for the actual business of producing dictionaries. He also believes that lexicographers should be encouraged to format their resources as structured data rather than in, say, Word or Excel.

Vera Hildenbrandt

Professional Background

Vera Hildenbrandt is Executive Director at the Trier Center for Digital Humanities, Trier University. She supervises projects at the interface between traditional and digital philologies, including digital lexicography (for instance, the digitization and digital publication of the first as well as the revised edition of the *German Dictionary* by Jacob Grimm and Wilhelm Grimm and the *Goethe Dictionary*). She is also a lecturer in the Digital Humanities Master of Science program at the University of Trier, currently teaching a seminar with exercises in digital lexicography. She is also part of a team working on the Trier Dictionary Network, a network of more than 20 different lexical resources.

In ELEXIS, she is involved in the conception of a modular Dictionary Viewer. The viewer will have a publishing module for dictionary entries, for the bibliography of its sources, for its preface and so on. The viewer will also include different visualization tools: for instance, a geographical visualization for dialect dictionaries, showing the spatial distribution of a word on a map, or timelines and so on. The user should be able to configure the viewer and its modules with a web browser.

Educational Background

After studying German and French Philology, Hildenbrandt wrote a dissertation in Modern German Literature on "Europe in Alfred Döblin's Amazon Trilogy. Diagnosis of a Sick Continent". At the same time, she worked as a research assistant in various projects at the Trier Center for Digital Humanities. Her studies included no formal lexicographic training but she was "already an eager dictionary user and lover" during her studies:

> *I am not a lexicographer in the proper sense, but rather a dictionary digitizer with a love for dictionaries and a good knowledge of the lexicographical process, dictionary typologies and dictionary structures. I developed my skills in **learning by doing**, analyzing dictionaries, reading research literature, and participating in workshops. I was also a member of different lexicographical networks.*

Hildenbrandt's most memorable educational experience when it comes to dictionaries was working on the first edition of the *German Dictionary* by Jacob and Wilhelm Grimm.

> *I've studied the history of the dictionary and the making of the articles, read what I could find in the research literature about it. I've done article analysis, talked to lexicographers who worked on it and on its revision. So I learned that dictionaries contain a whole cosmos that should be made accessible in a new way in the digital age.*

Skills Development

Retrodigitization is a complex endeavour, which requires familiarity with lexicographic processes, insights into article editing systems, knowledge of lexicographic macro and micro structures, knowledge of digitization processes, knowledge of markup languages and standards (XML, TEI, LMF, MDF), knowledge of usability and design, tools already available for digitizing and encoding dictionaries.

When teaching, Hildebrandt tries to combine theoretical and practical knowledge in seminars with hands-on exercises. The topics that she covers in her seminar -- and which, she believes are equally important in the context of ELEXIS -- are:

- Introduction to Lexicology and Lexicography
- Evaluation criteria for digital dictionaries
- Look what up? - An insight into dictionary types and typologies
- How are (digital) dictionaries created? - The lexicographic process
- How are digital dictionaries prepared for a variety of user actions? - Data Modeling: Text Encoding Initiative, Lexical Markup Framework (LMF), Toolbox and Multi-Dictionary Formatter (MDF)
- How are digital dictionaries designed? - An insight into the design and usability of digital dictionaries
- The user as author - possibilities of user participation
- How are digital dictionaries used? - An insight into dictionary usage research
- Why dictionaries? - An insight into the use of dictionaries in teaching and learning

In her seminar, she offers each learner a "dictionary godchild" which they work on during the semester.

Since Trier is mainly active in the field of retrodigitization, they offer in-house training courses in the field of digitization processes and scripting languages which can be used to process texts. In addition, they make it possible for their colleagues to take part in special training courses, for example in the field of the TEI.

The most challenging aspect of teaching new skills to students is helping them learn "how to think in a structured and analytic way." In the beginning, it also important to help students overcome the idea that dictionaries are boring.

Teaching Formats

Whatever the teaching format (face to face vs. online), it should be the right mixture of theoretical knowledge, practical work and entertainment:

> *In my experience, lexicographical theory is best taught through lectures with readable scripts. However, these lectures should not take up too much time and space. Learners should also be able to take things into their own hands, for example try out different digitization methods and tools. There should be instructive, but not too long, written instructions.*

Expert knowledge should be included, for example, via video tutorials. In addition, one could imagine a video about visiting "a dictionary factory".

Training Materials

ELEXIS bundles lexicographical and technical competences from all over Europe. The project should pass on these competences. Hildenbrandt would like to see central access to everything that happens in the field of digital lexicography. Where there is no training material yet, ELEXIS should develop and offer it and at the same time refer to existing training materials so that information is available in one place.

The training materials should be staggered according to different levels of education: absolute beginners, advanced learners, experts, so that everyone can immediately see where they should start.

Hildenbrandt would also like to see something like an "expert contact exchange" so that experts could be integrated either personally or via Skype etc. in teaching units at universities and the other way around.

Lara Hudaček

Professional Background

Lana Hudeček is a tenured Scientific Advisor at the Institute of Croatian Language and Linguistics. She is interested in a range of issues including language standardization, normativity and terminology. She is the author of a grammar and several books on language usage. From 2007-2013, she worked on the Dictionary of the Croatian Language for Schools. At the moment, she's working on the *Croatian Online Dictionary* (Hrvatski mrežni rječnik).

Educational Background

Hudeček studied Yugoslav Languages and Literatures at the University of Zagreb from 1979-1985. She obtained a Master's Degree in 1992 and a PhD in 2002 working on topics in historical linguistics. Initially, at the Institute, she worked on historical topics but then made the switch to contemporary language and lexicography. During her studies, she did not receive formal lexicographic training. She's a "**lexicographic self-learner**".

Her first lexicographic practice was when she worked on the book *Croatian Usage*, which came with a dictionary consisting of 80,000 lemmas. She was always "fascinated by dictionaries" and learned a lot by **consulting and analyzing existing dictionaries**.

Skills Development

According to Hudeček, **developing technological skills for lexicography is essential** in the digital age. Yet the main challenge for Croatian lexicographers is that there are **no systemic educational offerings** in this field. A single seminar on lexicology and lexicography at the university is far from sufficient especially considering that the curriculum is quite removed from contemporary trends and best practices.

Developing skills by attending various workshops is very useful, but "**the knowledge gained in unconnected workshops is partial and segmented**." What is missing is a complete, systematic, interdisciplinary curriculum:

> *In addition to building and collecting resources, ELEXIS should be a central hub for interdisciplinary lexicographic training and education.*

Hudeček believes that ELEXIS could play an important role in integrated curriculum development.

Teaching Formats

Hudeček **prefers face-to-face training measures** but is aware of the fact that organizing such measures may be costly and that ELEXIS will have to offer a combination of face-to-face and distant training measures.

Training Materials

Hudeček prefers following **written step-by-step instructions** than watching video, but she stressed that this a personal choice and that her younger colleagues may prefer watching videos. **Shorter videos are better than long videos.**

The main challenge for ELEXIS in the educational realm is that it will have to address very different communities, with distinct educational backgrounds and unequal levels of expertise. In order to make sure that the **ELEXIS training materials** can address the widest possible audience, they **should assume no previous knowledge whatsoever**:

> *Do not assume that anybody knows anything. Think how low you can go.*

As part of her work on the Croatian Online Dictionary, Hudeček developed a **glossary of e-lexicographic terms** (http://ihjj.hr/mreznik/page/pojmovnik/6/). This glossary "was a product of my own troubles" trying to understand a host of new terms as she was starting to work using digital tools in her day-to-day lexicographic work. The Glossary covers a wide range of tools and methods, but also

institutions and organization involved in e-Lexicography. It may be a good idea for ELEXIS to develop a similar kind of glossary.

Annette Klosa-Kückelhaus

Professional Background

Annette Klosa-Kückelhaus works at the department "Lexik" which is part of the Institute for the German Language (Leibniz-Institut für Deutsche Sprache, Mannheim). She heads the "Lexicography and Language Documentation" subdivision of "Lexik" and is involved full-time in lexicographic theory and practice. An important part of her work involves organizing different dictionary projects and facilitating the work of her colleagues. She is also a practical lexicographer, working on a German neologism dictionary and conducting research on such areas as internet lexicography, grammar and dictionary, word formation, and language norms.

Educational Background

Klosa-Kückelhaus has a Master's Degree (Magisterstudium) from the University of Munich, where she majored in German Linguistics with a minor in Archeology and History. Afterwards, she went to work in a publishing house, where she began her career as an editor for cook books before eventually returning to university and earning a PhD in German Linguistics on a topic of word formation in German. Subsequently, she took a job at the Duden publishing house. She developed her lexicographic skills through on-the-job training and has not attended any formal courses in lexicography.

Skills Development

Klosa-Kückelhaus believes there is no difference between the skills required for working on digital and and on print dictionaries. She argues that it is important to have a good feeling for language as well as an understanding of the commonly used technical tools. She thinks that it is desirable for potential lexicographers to have some level of technical understanding of lexicographic issues rather than just a purely linguistic training. She also stresses the importance of curiosity and fun working with language as well as perseverance and accuracy as desirable qualities in a lexicographer.

Klosa-Kückelhaus is involved in teaching for the European Master of Lexicography (EMLex). She uses textbooks, other publications on lexicography, lexicographic and corpus tools, and different dictionaries as materials to teach her students. She asks her interns to read what she regards as key publications in the domain, and to visit colleagues who are responsible for software development and technical support. She doesn't see any obstacles for developing skills in lexicography, instead she thinks that the problem is more often finding employment in the field afterwards.

Teaching Formats

Klosa-Kückelhaus suggests the use of online learning units, such as on the platform Moodle, although this may be expensive to set up and maintain. Webinars would also be useful because they enable educators to reach a large numbers of students at the same time and as well as allowing time for (virtual) interactions with students afterwards. Another suggestion is to have a combination of written materials, which can be discussed in a classroom/virtual settings, along with video tutorials dealing with

such topics as how to choose the best corpus examples to go in the entry. Annette thinks it would be extremely helpful to incorporate scenarios that involve lexicographers 'thinking aloud' and describing what they're doing as they carry out various tasks into lexicographic teaching.

Training Materials

ELEXIS could have an interesting role in terms of training lexicographers by offering materials for students who lack technical abilities but are good linguists. For instance, these technical skills aren't normally taught as part of the curriculum at German universities. She emphasizes the use of materials that are practical and focus on showing students/trainees how to do things; this is in the nature of the subject since lexicography is very practical and it would not be that useful to have an abundance of theoretical training material.

Klosa-Kückelhaus believes that it is important to convey the pleasures of lexicography in everything that educators in the field of lexicography do: in the teaching and training that they carry out and in the materials that they produce. At the same time, it is also necessary to encourage students to appreciate how much drudge work can be involved. In her experience, interns and students sometimes come to the field with a mistaken view of it and how repetitive it can be. A well-balanced set of training materials will communicate both aspects of lexicographic practice and will assist interns in understanding if lexicography is the kind of career they want to go into.

Nikolče Mickoski

Professional Background

Nikolče Mickoski is a research associate at the Lexicographic Center of the Macedonian Academy of Sciences and Arts. As part of his job, he works with professionals in different fields, preparing dictionaries that contain specialized and professional terminology to be used by content creators, translators and other professionals. The institutional project he's affiliated with -- *Macedonian Scientific and Professional Terminology* -- is one of the first projects of the Academy, which started back in 1968 with the goal of compiling multilingual dictionaries for different terminology areas such as medicine, agriculture, economy, technology, theology, engineering etc. One of his professional goals is to digitize over 40 such dictionaries that currently exist only in hard copy and make them available online.

Educational Background

Mickoski is a professional interpreter by training: he studied translation and interpretation at the Faculty of Philology at St. Cyril and Methodius University in Skopje, eventually obtaining a Master's degree. He is currently pursuing a PhD in the terminology and terminology management with a focus on the terminology of information society.

His working languages are Macedonian, English, German and Serbian. His interest in terminology and lexicography developed while he was working for various high-profile professional clients such as Microsoft and Sony on localizing their products for the Macedonian market, but his training included no specific lexicographic coursework but he has taken every opportunity to expand his knowledge by attending various training measures such as an advance online course to become a certified terminology manager in 2014, covering topics ranging from the quality assurance of terminological

entires to terminology standards and legal issues; a CLARIN workshop "Translation memories, corpora, termbases: Bridges between translation studies and research infrastructures" in 2018, which was "great for networking and for getting a deeper knowledge of the lexicographic field"; and, also in 2018, the DARIAH-ELEXIS Lexical Data Masterclass, where he learned a lot about TEI and GROBID dictionaries.

Mickoski's most memorable learning experience was working hands-on with clients defining workflows, preparing glossaries, working iteratively: "Before that I was learning more theory, and when I had to implement everything into practice, I learned a lot in the entire process working with a big client who guided me, especially in terms of quality assurance." Lexicographic practice always trumps lexicographic theory: "No matter how much lexicographic theory you learn, it doesn't help unless you actually start producing dictionaries."

Skills Development

Mickoski believes that the most essential skill for lexicographers to learn these days is how to use corpora in their day-to-day work. This is especially difficult in the context of North Macedonia because of the lack of a balanced, national corpus, the lack of well-trained POS taggers for annotating corpora, and the lack of a corpus linguistic curriculum at universities: "We should strive to teach students how to use technology in their everyday work."

The challenge is both technological and social. The lexicographic landscape in Macedonia is rather conservative:

> *Terminologists and dictionary makers work on their terms and words which are chosen arbitrarily according to their preference and they do not provide additional justification about why a certain term was included in a glossary or in a dictionary.*

Because they do not have the necessary resources which would provide statistical data about usage, dictionaries often include words or translation equivalents that their authors think are the most suitable but may entirely miss alternatives which may already be in use.

According to Mickoski, it is also important for lexicographers to be familiar with intellectual property rights and copyright issues regarding legacy dictionaries: "we should use only resources that are available to us without stealing content."

Finally, lexicographers need to learn how to present lexical resources online and make them available to all the users, no matter where they come from: "A dictionary which is printed in only 300 or 500 copies is a useless dictionary because only a limited number of people can use it."

ELEXIS, according to Mickoski, "should be a hub, a central point where lexicographers can go and get updates about new software, new technologies, new developments, new conferences, trainings and everything."

In order to help users develop skills, ELEXIS should try to match more advanced users with those from lesser-resourced communities:

> *If we speak concretely about Macedonia, for instance, it would be great if colleagues who work on similar languages such as Serbian or Bulgarian could organize a workshop on building corpora and training POS taggers so that we can learn how to do this ourselves.*

In terms of Travel Grants which are offered by ELEXIS, Mickoski said he couldn't find "concrete information" about what he could learn at each institution.

Teaching Formats

Nickoski has plenty of experience participating in both online and face-to-face measures, and he has no doubt about which kind of training he prefers:

> *I can compare various online courses on with recorded videos and homework assignments and the DARIAH-ELEXIS Lexical Data Masterclass which was conducted face to face in Berlin. Face to face training is the best. In the five days I spent in Berlin, I managed to digitize a dictionary from scratch without any previous knowledge of GROBID dictionaries.*

Short-term training measures (workshops, masterclasses) in which "participants develop skills in hands-on sessions through the process of making and correcting mistakes" are better suited for learning specific and more advanced skills, according to Mickoski. More general skills -- or more complex processes such as creating a dictionary from scratch -- should be taught in university courses or over longer periods of time.

Training Materials

Mickoski likes instructional videos or screencasts because he considers himself a visual learner. The problem with online training materials in general is that often there is no one to turn to in case users get stuck:

> *It would be great if those videos are accompanied by a transcript and some kind of support: where students can ask a question when they have a problem.*

Written step-by-step instructions are easier to follow because, for instance, you can copy and paste certain commands, which you can't do from a video. But step-by-step video instructions have one advantage over written training materials because they tend to *show* rather than *describe*, and showing is more complete:

> *The author of written step-by-step instructions may omit some steps because he or she thinks that the participant already knows them.*

Ideally, according to Mickoski, videos should come with a full transcript or be combined with written instructions.

Emrah Özcan

Professional Background

Emrah Özcan is a research assistant at Yildiz Technical University in the English Language Teaching Department and a PhD student at Ankara University's Linguistics Department, where he's he's writing a dissertation on noun formations as they are represented in Turkish dictionaries. In addition to his dissertation, he's currently involved in a project studying the changes in Turkish vocabulary as evidenced by the four different editions of the Turkish Dictionary published by the Turkish Language Association since 1993.

Educational Background

Özcan has an undergraduate degree in English Language Teaching Istanbul University, and a Master's Degree in Turkish as a Foreign Language from Yildiz Technical University. For his Master's thesis, he designed a prototype of a semi-automatically generated Multilingual Learner's Dictionary of Turkish. In his studies, Özcan did not receive explicit lexicographic training. He considers himself a **self-learner**: "I read books and learned from the internet: I basically trained myself."

He attended the DARIAH Lexical Data Masterclass in 2017 and the DARIAH-ELEXIS Masterclass in 2018, both of which he describes as "milestones" in his scholarly career as far as retrodigitization and digital humanities are concerned. Dictionaries, however, have been an important part of his education all along.

> *I used to look up every single word I didn't know. And sometimes I would look up the words from the definitions I was reading. My friends who would skip the dictionary part would often finish their reading before me. It was something like a journey for me because with each word, each entry, I felt I was going to a different destination.*

Another memorable experience for Özcan was when he discovered a misspelled word in a Cambridge University Press dictionary:

> *My teacher didn't just say oh well, good job, but she informed representatives of CUP in Turkey. They came to our school and gave me a bunch of presents (including grammar books and readers) in a ceremony in front of my entire class.*

Skills Development

According to Özcan, **lexicographic and technological skills** are mutually complementary and **need to be developed in tandem**. One of the side-effects of learning how to model lexicographic data in XML is that scholars get to know and understand their dictionaries better: "This is essential," said Özcan "[Data modeling] does not only help you understand what's inside your dictionary, but also how it is constructed or how it evolves over time, through different editions". Dictionaries reflect social and political changes: for instance, after 1980, the year of the military coup, the Turkish Dictionary included more words of Arabic and Persian origin, echoing the conservative and religious turn of the society at the time.

Scholars who do not learn how to apply technological skills to studying dictionaries can achieve "limited results", according to Özcan. But digital technologies are important not only in the study of old dictionaries. Technology makes it possible to include new features in modern-day dictionaries and help users along the way: "When you look up a word in the Collins dictionary and you see a little mark which says that this word is within the first 1000 of the most commonly used words in English, then you realize that this is a very frequent words and that you should probably learn it."

One of the main challenges facing dictionary research, according to Özcan, is the **lack of open-access access to lexical datasets**. In order to study the lemma lists of different editions of the published dictionaries, Özcan had to learn how to retrodigitize existing dictionaries, combining OCR, GROBID Dictionaries and text encoding in TEI. The data of the Turkish Dictionary is not open even though the Turkish Language Association is a public entity, which, according to Özcan, is "not a best-practice example, considering that we live in the digital age."

Another problem that has been singled out by Özcan is the **lack of established university-level courses in lexicography**: "There is no master's or PhD course in lexicography in Turkey that I'm aware of." Lexicographers in Turkey "learn on the job."

According to Özcan, most **lexicographers are not familiar with XML and XSLT**. Furthermore, these **skills are often perceived as too technical**: "When I say XML or XSL or TEI, they think that it is only for the computer science people."

In his research on the Turkish Dictionary by the Turkish Language Association, Özcan found out that some of the previous editions are stored in databases, but none of this information is readily available in XML. There is **no culture of sharing lexicographic data**: "This is in my opinion one of the main obstacles for Turkish lexicography: we do not want to share, we just want to print our dictionaries and that's it." Özcan attributed some of this reluctance to sharing to **fear regarding the safety of digital resources**, the possible **data abuse** as well as the **financial factor**: "People are worried that their data could be stolen, that it could be used without their consent and that they will lose money."

Academic culture is built on the **"publish first" principle** which is understandable and linked to the way academics get credit for their work. But "being open and sharing is the only way to go now," said Özcan.

Teaching Formats

According to Özcan, ELEXIS should **provide as much face-to-face training as possible**: "When you bring the right people together, like you did in the Lexical Data Masterclass, you can achieve more in a week than you would in month's time on your own." Face-to-face training opportunities are not only intense, but participants are given a time and space for learning in which they are not "distracted by their day-to-day obligations."

ELEXIS could, according to Özcan, specifically **target participants from lesser-resourced countries** in order to help overcome the digital gap in lexicography, "as a kind of positive discrimination."

According to Özcan, online training measures such as video tutorials work for basic and general instruction but "**solving specific problems requires face-to-face training**." This may not always be possible to organize, due to financial and other constraints, but perhaps one could **create a forum**: "Whenever you come across a problem, it's good that you can find some other people who already had encountered the same problem."

Training Materials

Still, **videos are useful for showing step-by-step instructions**: "I learn greatly by seeing how something gets done, step by step". The choice of video vs. written step-by-step instructions is not a matter of pedagogical efficiency but rather a personal preference based on one's learning style: "I'm a visual learner so I think I would prefer to watch step-by-step videos."

Videos have the added advantage that they can contain **subtitles**: the text of the whole video "can also be extracted as a kind of script to follow." Özcan prefers **short, to-the point videos** in which he can learn things that he can test immediately: "just give me a couple of things that I can start with and **see results** so that I will **feel motivated to continue**."

But even in written tutorials, screenshots and/or embedded videos are essential.

Ideally, training materials should, according to Özcan, be accompanied by a forum (see above) and a "feedback section" where users can send their ideas and suggestions about the topics to be covered.

Creating communities around training is good for both the communities and the authors of the training materials: responsive members of the community can help each other, while the authors can get feedback on their training materials.

Laurent Romary

Professional Background

Laurent Romary is directeur de recherche (director of research) at INRIA, a French national science and technology institute dedicated to computational sciences, and former director of the Digital Research Infrastructure for the Arts and Humanities (DARIAH). His research interests include various types of data modeling and information extraction issues in the humanities at large. Romary describes himself as a "pure computer scientist" who "switched to computational linguistics at the very early stage of [his] career." He has supervised for many years a research team consisting of both linguists and computer scientists working on language resources and has been involved in a number of activities related to the creation of language resources in general, and lexical resources, in particular.

Romary is also very well known for his work on standardization. He has been involved with the Text Encoding Initiative (TEI), a de-facto standard for encoding humanistic texts, since 1992, serving over the years as a member of the TEI Technical Council (2001-2011) and its chair (2008-2011). In the context of TEI, Romary served as one of the early proofreaders of the TEI Dictionary Chapters and is nowadays working in the context of the DARIAH Working Group "Lexical Resources" on TEI Lex-0, a customization of the TEI schema that serves as a baseline encoding and interchange format for dictionary data.

He has also been active within ISO, as chair of ISO/TC37/SC4 (2002-2014) and chair of ISO/TC37 (since 2016).

Educational Background

Romary received a PhD degree in computational linguistics in 1989 and his Habilitation in 1999.

Romary emphasized that from an early stage of his education he was trained in databases, tree structures, graph structures and other types of representation, which made him recognize the value of "having models like the one we have in the XML domain that are readable by machines but also understandable by users as opposed to complex graphs for instance."

Even though he didn't receive formal lexicographic training, he was very much influenced by his involvement with large dictionary projects such as the *Trésor de la langue française*, which he converted from an early proprietary format to a "decent TEI representation," and also in terminology with the *International Hydrographic Dictionary*, for instance, which he computerized in the context of the European Research Project *Dhydro*.

Throughout his career, Romary worked closely with linguists and lexicographers:

> *It can be very dangerous if computer scientists pretend that they are lexicographers or that they know about lexical data without working hand in hand with linguists or lexicographers.*

For Romary, the most memorable educational experience in lexicography was not formal but arose out of a close collaboration with John Sinclair:

> *For a couple of years, I worked quite closely with John Sinclair, one of the initiators of the COBUILD Dictionary. We were thinking a lot about what a good corpus for building a dictionary could be, what's the use of encoding, since John was pretty much against any kind of tagging and encoding of corpus data... There were a lot of debates there, we would not agree on everything but at the same time we recognized what each of us could bring to the debate.*

Skills Development

According to Romary, the most essential skills required for working with lexical data are the ability "to abstract the information structures across a large number of dictionary entries" and "to differentiate between representation and presentation", i.e. between the abstract model and its actual typographic realization on paper or computer screen.

> I've noticed that lexicographers without any modeling background sometimes see each lexical entry as something very different from all the others and have difficulty abstracting away. This is a real skill.

According to Romary, the biggest challenge we face in terms of helping people develop new skills is reaching out to new audiences and incorporating technical skills into university curricula:

> *Our experience with the Lexical Data Masterclass is that there is a real expectation from the community that we do such things. The participants are very happy, and I think these types of events contribute a lot to the defragmentation of the landscape in terms of representation because a lot of participants are actually switching from proprietary formats to something which is more compliant like with the TEI Guidelines.*
>
> *So the the challenge is now probably to make sure that these kind of skills are not just targeted at experienced lexicographers (although we had quite a few of them in the masterclass) but to ensure that they are integrated in curricula, for example at master's level, for instance. So that people right from the beginning doing linguistic studies or lexicographic studies would also be trained in those technical practices.*

According to Romary, the role of a research infrastructure in this context is "to stabilize the knowledge." This can be achieved by creating training materials, documenting training measures, publishing blog posts, sharing samples online. In addition, the infrastructure should contribute to the creation of basic, introductory documents. Such introductory documents can also help with face-to-face training because participants can already prepare ahead of time and start working on their own samples sooner rather than later. The collection and creation of such training materials is time-consuming but the added advantage of investing time in training measures -- including face-to-face training -- is that infrastructural services can benefit from being tested in face-to-face settings:

> We see a lot of different real-life examples [when we conduct training measures] so we can make sure that what we put in TEI Lex-0 reflects the expectations and constraints of the projects we are dealing with in the classes we teach.

Teaching Formats

Romary is a believer in **face-to-face training measures** and **practical, hands-on teaching methods**: "The more I teach, the less I shows slides and the more I directly do things with students."

It is in this context that the idea of the Lexical Data Masterclass was developed:

> The Lexical Data Masterclass is a kind of place where participants come with their own projects and issues, but when you put them around the table you discover that they ask themselves the same type of questions. So there is a real dynamics of working with the data and trying to find real solutions together: as soon as they do a thing once on their own, you know that they've mastered the concept, that they know how to combine an example with a bibliographic reference, where to put the translation, what are the issues with the various fields in a technological description and so on. So this is really the kind of live activity which epitomized the face-to-face work.

When teaching, Romary says he usually starts by making quite a **theoretical introduction concerning lexical models** and describing the difference between **onomasiological and semasiological representation**; quickly followed by discussing actual examples and doing things completely with XML, introducing the *TEI Guidelines* at an early stage.

Romary values **iterative encoding exercises**: starting with simple representations and moving from there to **gradual enrichment with additional sets of information**.

> I've discovered that basically you cannot teach data modeling especially in the domain of lexical information if the seminar participants do not have their own projects or their own data on which they can practice.

Training Materials

Ideal training materials, according to Romary, **provide step-by-step instructions** with **various degrees of difficulty**: "with appropriate annotations so that people can be guided through the process, following a script." This is something Romary does in vivo when he teaches lexical data modeling, but the difficulty with producing such instructions to be used asynchronously is again that they take a long time to create.

> *I dream of having an Egyptian scribe who would transform the face-to-face instructions I give into some kind of training material.*

Videos on their own may not be enough. At first glance, videos can be useful for "getting acquainted with the material," but it would be even better if videos were accompanied by full-text transcription or embedded in some kind of textual format "so that afterwards you can look at the outline rather than the video itself again."

Videos can be useful, especially for visual learners, but **the challenge with videos** stems from the need to find **the appropriate length** and provide **enough contextualization**.

> *I don't know if there are really many people who would follow a MOOC on representing lexical data in XML. I wouldn't have the patience myself.*

Taking the example of TEI Lex-0, Romary suggests that videos could be created but only if the entire domain is split into a sufficient number of small units so that you could, for instance, have a separate video on how to encode grammatical information, or a video on encoding etymons and etymological descriptions, etc.

Ana Salgado

Professional Background

Ana Salgado is a lexicographer with 16 years of experience, currently working at the Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLLP) of the Academia das Ciências de Lisboa in Portugal. Previously she worked at a large Portugugese publishing house Porto Editora. Salgado's research interests are in the areas of linguistics, lexicography and digital humanities with a focus on applying digital humanities methods in the area of Lexicography.

As the coordinator of the Portuguese Academy Dictionary, Salgado continue to practice the daily tasks of a lexicographer: research into and identification of new words that are in common usage for inclusion in the dictionary; reviewing and editing existing definitions; checking the corpus for evidence of meanings and usages of words or multiword expressions etc.

Educational Background

Salgado studied Portuguese at the Faculdade de Letras da Universidade do Porto. She didn't plan on becoming a lexicographer even though she has always had a great passion for languages.

> *My parents had an excellent collection of dictionaries. A dictionary was like the Bible to me. Only later did I understand that it was not quite like the Bible.*

It just happened that her first job was in Porto Editora, in the Department of Dictionaries, working on Spanish dictionaries. She was part of the team that developed the morphological analyzer, then the spelling converter, a mobile application, and was eventually entrusted with the scientific coordination of the department. Before that, her only introduction to the subject of lexicography was due to the fact that a professor of hers, Mário Vilela, was working on a usage dictionary. But there was no specialization in lexicography at her university.

Although she had no formal lexicographic training at the time, she read a lot, attended conferences, studied and compared different Portuguese, Brazillian and English dictionaries.

Salgado is currently a PhD student at the Centro de Linguística da Universidade NOVA de Lisboa (CLUNL). She's working on the treatment of terms in general language dictionaries. The combination of lexicographical and terminological methodologies can be an added value for the organization and description of lexicographic articles and improving the quality of lexical databases. She entered the PhD program in order to "deepen her knowledge of theory and to substantiate her know-how." This, she feels, was necessary because "there is still a large gap between the industry and academia."

Skills Development

For Salgado, it is important to develop a range of skills : linguistic, lexicographic (macrostructure/microstructure) and technical. When it comes to technical skill, it is especially important to be familiar with data modeling best practices: it is important to know XML and how to work with an XML Editor such as oXygen.

For people working on retrodigitizing dictionaries, the knowledge of the *TEI Guidelines* is essential.

> *Lexicographers who worked for many years on printed dictionaries, must make an effort to break with the past and embrace the era of digital revolution in order to learn how to benefit from it. We need to* ***broaden the scope of dictionaries from a product that fulfills a single purpose****. Through the usage of lexical organized databases, we are able to extract larger and more diverse information and* ***build new products that can satisfy multiple user needs****. This is our biggest challenge*.

At the same time, Salgado is aware of the fact that some lexicographers are reluctant about working with specialized editors or learning standards: "They keep using Microsoft Word to write dictionaries. That's a big problem".

The challenge that traditional lexicographers face is not only technological but social and cultural: "we have to change the perspective."

Teaching Formats

ELEXIS already plays an important role in skills development: Salgado received an ELEXIS travel grant to visit the Real Academia Española and she attended in the DARIAH-ELEXIS Lexical Data Masterclass. At the RAE, she got to know the infrastructure of the *Diccionario de la lengua española*, and at the masterclass, she worked on expanding her knowledge of TEI and revising the encoding guidelines of the Portuguese Academy Dictionary according to TEI Lex-0. She appreciated both measures as excellent learning opportunities.

> *I have a feeling that we were working in isolation before. For example, between Portuguese and Spanish, there are so many similarities that I came back from Madrid with a dream of an Iberian agreement. If we combined efforts, it would be ideal. For the sake of interoperability, it is so important to know what technologies we are using, if they are mutually compatible and which are the best standards. I think that is also the aim of ELEXIS.*

More workshops and training opportunities in the context of ELEXIS would be welcome, according to Salgado.

Training Materials

Salgado preferrs step-by-step written tutorials and hands-on exercise.

The main problem with training measures and training materials, however, is that they are limited in time and scope and usually not focused on continued community-building:

> *It's great to have a training week, but after the training is over, the students are alone again and without support.*

ELEXIS should think of how to create a **continuous support mechanism** beyond individual training measures or research visits. Salgado would “love to contribute more, but it is not clear to me how I could do it.”

One idea that comes to mind is **creating a blog with exercises to be completed after the training** so that the community would continue to interact.

Appendix I: Interview Script

Introduction

Let me tell you a little bit about this interview and the context in which it is taking place. As part of the ELEXIS WP5 on Training and Education, we're currently conducting interviews with lexicographers, both from within our Consortium and outside, in order to get a better sense of the kinds of skills that are needed in order to work competently as lexicographers in this day and age and to better understand the kind of training resources that are needed for the development of those skills.

These qualitative interviews will complement the quantitative surveys that have also been conducted by ELEXIS.

We are speaking to a wide range of colleagues: some are working on digital-born lexica, some are working on regrodigitzed lexica, some come from well-resourced and some from lesser-resourced languages and communities.

This interview will be recorded in order to help us write the report, but the recording itself will be kept confidential and the video will not be released. This is also indicated in the consent form which we've asked you to sign.

The report that we write will contain a section on each interviewee. The section will provide basic information about you, your institution and your position. And it will be followed by an analytic summary of the interview.

Do you have any questions before we start?

Questions

1. Please state your full name, your current position and your institution.
2. Could you please describe briefly what are the main aspects of your job and/or research interests.
3. To what extent are you currently involved in lexicographic theory and practice?
4. Could you please tell me a little bit about your own training and education: where did you study and what?
5. Did your studies include specific lexicographic training?
    a. if yes, in what form? (university courses, workshops, on-the-job training for specific projects etc.)
    b. if not, how did you develop your own lexicographic skills?
6. If you look back at your career, what was the single most memorable educational experience for you when it comes to lexicographic theory and practice? Where did you learn the most? And how?
7. What skills do you consider essential in your field?
8. How do your students/new colleagues develop those skills?
    a. if you are at a teaching institution, how do you do it? what kind of courses do you offer?
    b. if you are not at a teaching institution, how do you or how does your institution help new colleagues when they enter the job?
9. Does your institution invest in continued skills development of your staff?
    a. if yes, in what way?
    b. if not, what would be in your opinion the best way of going about it?

10. What are in your opinion the main challenges and obstacles to the development of lexicographic skills?
11. What do you think should be the role of a research infrastructure like ELEXIS in helping lexicographers develop new skills and overcome those obstacles that you mentioned?
12. What would be in your opinion an ideal teaching format for developing new skills?
13. Training materials come in different shapes and forms. Please describe what you would consider to be your ideal training materials.
14. In ELEXIS, we'll be developing our own training materials. Could you give us any advice on what to do or not to do?

Conclusion

Thank you very much for taking the time to speak to us. Is there anything else in the end that you would like to share with us that hasn't been covered by the previous questions?

Appendix II: Consent Form

I consent to the processing and transmission of my personal data, specifically my name, position, institution, educational background and the views I expressed during the interview for the ELEXIS Skillset Report. I understand that the said report is being compiled as a public deliverable in the H2020-funded project European Lexicographic Infrastructure (H2020-INFRAIA-2016-2017, Grant Agreement No. 731015) by members of the Work Package 5 "Training and Education."

I consent to the video recording of my interview being stored for up to four years and accessible only to the authors of the said report. The video recording shall not be publicly released in part on in full.

This consent can be withdrawn at any time without explanation by emailing WP5 leader TomaTasovac at ttasovac@humanistika.org. Withdrawing consent does not affect the legality of earlier processing.

_____

Name (in print) and signature