Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs

Jelena Kallas[1], Svetla Koeva[2], Margit Langemets[1], Carole Tiberius[3], Iztok Kosem[4]

[1] Institute of the Estonian Language, jelena.kallas@eki.ee, margit.langements@eki.ee

[2] Institute for Bulgarian Language, Bulgarian Academy of Sciences, svetla@dcl.bas.bg

[3] The Dutch Language Institute, carole.tiberius@ivdnt.org

[4] Jožef Stefan Institute, iztok.kosem@ijs.si

Abstract

The paper presents the results of a survey on lexicographic practices and lexicographers' needs across Europe (and beyond) both for born-digital and retrodigitized resources. The survey was conducted during the period from 11 July to 1 October 2018 in the context of the Horizon 2020 project ELEXIS (European Lexicographic Infrastructure). The survey was completed by 159 respondents from a total of 45 countries, comprising 36 European countries and nine countries outside Europe.

Looking in detail at the results of the survey, the paper focusses on determining what constitutes a job description of a modern lexicographer, including the training needed. One of more notable findings is that lexicographic training is still in most cases provided by the employer rather than obtained through formal education programmes. Furthermore, a list of various dictionary-writing systems and corpus-query systems is provided, including their features currently most often used by lexicographers. Accompanying this is information about the features lexicographer want or need in their tools. Also, the paper offers insights into current trends in lexicography and what lexicographers see as the most important emerging trends that will affect lexicography in the future. Overall, these results provide a detailed insight into what is needed in terms of tools and training and thus feed back into the ELEXIS project and will help to fine-tune resources within ELEXIS.

Keywords: e-lexicography; lexicographers' needs; survey; lexicographic practices

1. Introduction

In lexicography, there is a lot of research available on methods of dictionary compilation, dictionaries, and dictionary users and their needs. On the other hand, until recently at least, there has been little literature on lexicographers, their practices and needs. This has become even more important in the age of significant changes in lexicography, brought about by technological progress and the move from the print medium to the digital one. The need to bring lexicographers from different countries together in order to tackle the challenges of modern-day dictionary making has thus become even greater.

The first steps towards addressing this issue were made in the European Network of e-

Lexicography (ENeL)[1], a COST Action funded by the European Union that brought together nearly 300 lexicographers from 30 different countries. Other than enabling the exchange of knowledge and expertise, ENeL produced highly valuable results such as various surveys among lexicographers and their institutions and a European survey on dictionary use (Kosem et al., 2019), the largest dictionary user survey to date.

One of the most important outcomes of ENeL has been ELEXIS (European Lexicographic Infrastructure)[2], a Horizon 2020 infrastructure project dedicated to lexicography. This new infrastructure aims to enable efficient access to high quality lexicographic data, and to bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources. ELEXIS activities have used the results of ENeL, however further research was needed to obtain a detailed insight into current lexicographic practices and the needs of lexicographers. Consequently, two surveys have been carried out within ELEXIS focussing on various aspects of the lexicographic workflow such as software and tools, publication, retrodigitization, metadata and data formats. The first survey was targeted specifically at individual lexicographers. The second survey focussed on institutions and targeted senior lexicographers and IT specialists from eleven ELEXIS lexicographic partner institutions. The survey for institutions is not part of this paper (the results are presented in detail in ELEXIS Deliverable 1.1, see Kallas et al., 2019), however we include relevant findings from the survey when appropriate.

In this paper we initially provide the background of the survey. Then, in Section 3, we introduce the general principles, aims, structure and the implementation of the survey, followed by the presentation of the results in Section 4. The last two sections are dedicated to the discussion on the implications of the survey findings for lexicographic practices and ELEXIS efforts, and conclusions about the overall value of the survey and future plans in the ELEXIS project.

2. Background

The information on lexicographic practices has often been generalized based on selected project(s) or researchers' experience (cf. Hartmann, 2003; Atkins & Rundell, 2008; Klosa 2013). While such works are very important for lexicography and manage to show the state-of-the-art of the discipline, they do not point out the differences and similarities between lexicographic practices in different countries or even at different institutions.

One of the main projects/initiatives that helped gather a great deal of information on lexicographic practices across Europe, and thus fill some of these information gaps, was

[1] http://www.elexicography.eu/ (1 July 2019)

[2] https://elex.is (1 July 2019)

the COST action European Network of e-Lexicography (ENeL). Within the Action, a number of surveys have been carried out, and we summarise the most relevant here.

The first survey, conducted in 2014, focused on the workflow of corpus-based lexicography. Six general monolingual dictionaries, one bilingual dictionary, and seven specialized dictionaries and databases were covered. All 14 resources were published online, one also in print (at the time). The main findings stated in the report (Tiberius & Krek, 2014) were that the role of computers in lexicography is continuously increasing, but the compilation of dictionaries is still a highly labour-intensive task. Most projects followed to a certain extent the phases of the lexicographical process proposed by Klosa (2013), with the analysis phase taking by far the most time. Lack of IT support was one of the problems mentioned by the majority of projects. Some attention was also paid to user involvement, with the main finding being that users need to be involved in the later stages of a lexicographical project (afterlife, etc.); crowdsourcing was mentioned as one option of earlier involvement, but it was concluded that more research (and a separate survey) was needed on this subject.

The second survey, conducted in 2014/2015, focused on Dictionary Writing Systems (DWS) and Corpus Query Systems (CQS) (Krek et al., 2014). It consisted of 94 questions and was completed by 69 lexicographers and computational experts (computational linguists, software developers, etc.) from 35 different institutions in 25 different countries. The part of the report dealing with DWSs showed that 10 institutions used off-the-shelf products, 12 institutions developed their own software, whereas 16 institutions used customized software (XML editors, databases, wikis, etc.). In terms of functionality, most DWSs supported validation and consistency checking, and offered the use of templates for common dictionary structures. On the other hand, many DWSs did not include a spellchecker or integration with a CQS. As far as CQSs were concerned, 65% of the respondents reported that their institutions used them – by far the most widely used was (no)Sketch Engine[3] (11 institutions), followed by IMS Open Corpus WorkBench[4] (4). The evident trend was that open-source and commercial CQS at the time met the needs of lexicographic projects, while this was not the case for DWS considering the share of institutions that developed or had been developing in-house solutions.

The third survey, also conducted in 2014/2015, dealt with automatic knowledge acquisition for lexicography (Tiberius et al., 2015). It consisted of 134 questions and was completed by 51 respondents (lexicographers, software developers, computational linguists, etc.) from 20 different countries. Thirteen different types of lexicographic data were proposed on the list of data types that could be automatically acquired from a CQS. The results revealed that more commonly extracted types of lexicographic data

[3] https://www.sketchengine.eu/nosketch-engine/ (1 June 2019)

[4] http://cwb.sourceforge.net/ (1 June 2019)

were lemma lists, frequency information, example sentences, grammatical patterns, and multiword expressions (ranging from collocations to idioms), while other types such as form variants, neologisms, translation equivalents, lexical semantic relations, word senses, linguistic labels, definitions and Knowledge-Rich Contexts[5] were automatically extracted by only a few institutions. Given the state-of-the-art of lexicography at the time, it is slightly surprising that, for example, translation equivalents, lexical semantic relations (e.g. synonyms) and linguistic labels were automatically extracted only by a few institutions. This finding is particularly interesting in the case of translation equivalents and lexical semantic relations, as they were reported, after lemma lists, frequency information and example sentences, to be among the types of data that were integrated in the published dictionaries without intervention.

The aforementioned three surveys (Tiberius & Krek, 2014; Krek et al., 2014; Tiberius et al., 2015) provided a great deal of insight into lexicographic practices around Europe. Still, in some cases requesting more elaborate answers from the respondents should perhaps have given better results. Moreover, it would be better if all the surveys were conflated into one so that a more general picture per institution or respondent could be obtained, and that the questions could be connected. Finally, the number of institutions, and to a lesser extent countries, could be greater.

The survey conducted in the ELEXIS project and presented in this paper aimed to address some of these shortcomings. Also, due to rapid changes in lexicography and related disciplines an update to this overview of lexicographic practices was very much needed. One aspect that we wanted to add to this overview was the education and training of lexicographers, and their needs related to this.

3. Survey of Lexicographers' Needs

3.1 General principles and aims

The main aim of the survey was to get a good overview of lexicographic practices across Europe both for born-digital and retrodigitized resources, different tools and methods used by lexicographers around Europe, as well as the needs that they have now or anticipate to have in the short-term and long-term future. However, the survey was also disseminated outside Europe, as we were also interested in lexicographic practices around the world. In order to get as many responses as possible, we limited the length of the survey.

Many different channels were used for disseminating the survey, e.g. international and national mailing lists, social networks (e.g. ELEXIS Facebook and Twitter profiles),

[5] In terminography, a sort of hybrid of a good example and a definition, illustrating the meaning characteristics of a term, but not being a formal definition.

group or individual emails, a booth at the EURALEX conference, etc. It was important to get a good coverage of countries to enable comparisons, and more importantly, to help us in preparing more targeted activities with the ELEXIS project, such as training workshops and materials, and help to fine-tune resources developed within the project.

Equally important was the attempt to get several respondents from the same country, in terms of institution, age, role in the team, dictionary project, etc. to ensure that the data would be representative of a country and not of a single institution, generation, project and so forth. As a result we managed to obtain answers from a rather heterogeneous group of respondents in terms of their experience, employment status, projects they are involved in (types of dictionaries, language etc.), and the country in which they are based (see Section 4.1). This to some extent ensures that the results can be generalized to the lexicographic community as a whole.

3.2 Structure and implementation

The method chosen for the survey was an online questionnaire. Several survey tools were considered, and in the end Google Forms[6] was chosen as it is simple to use and manage, and it covered the majority of our needs. The survey was publicly announced on various mailing lists on 13 July 2018 and was closed on 1 October 2018.

The survey [7] contained 44 questions divided into six sections, i.e. (1) General information; (2) Ongoing work; (3) Software and tools; (4) Publication; (5) Retrodigitization; (6) Past and future. There were three different types of questions used in the survey: (1) "yes/no" questions, (2) multiple choice questions, and (3) open-ended questions. Not all questions were obligatory.

4. Results

4.1 Respondents' background and projects

The survey was completed by 159[8] respondents from a total of 45 countries, comprising of 36 European countries (140 respondents) and nine countries outside Europe (19 respondents). We decided to categorise under European countries those nations with close cultural ties to Europe (and inclusive status in EU-funded initiatives such as

[6] https://www.google.com/forms/about/ (1 July 2019)

[7] The survey for institutions was more detailed, containing 86 questions but divided into the same six sections. Both questionnaires can be found in the appendix of ELEXIS Deliverable 1.1, see Kallas et al., 2019). Because of data privacy issues the raw data cannot be shared.

[8] As some questions were optional, not all questions were answered by each respondent. For this reason, we provide the number of responses for each question (i.e. N = number_of_responses) in the results.

COST Actions) and with active partners in the ELEXIS consortium.

Figure 1 illustrates that the majority of the respondents work as full-time or part-time in-house employees and less than one quarter as freelancers (mainly the respondents from Northern Europe and the USA).

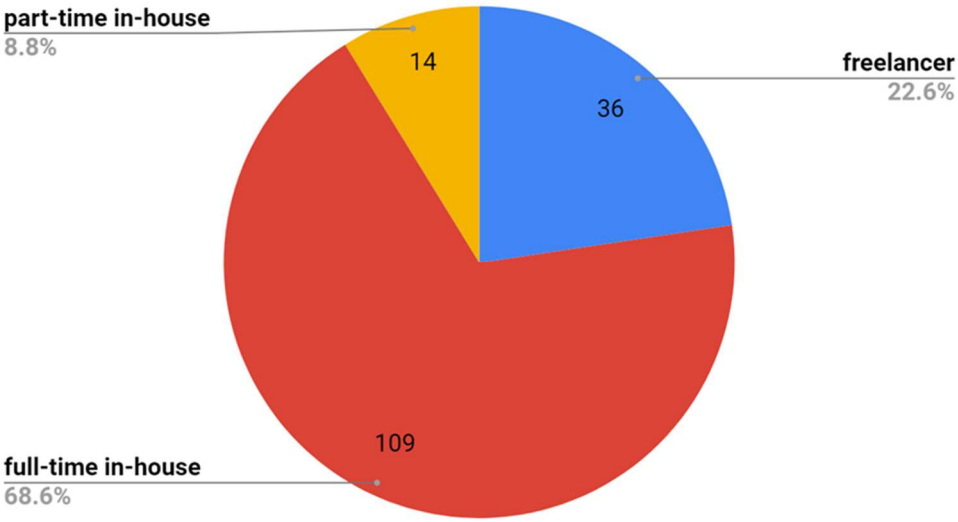

Figure 1: Employment (N=159)

A total of 77.9% of in-house lexicographers (123 respondents) work at public institutions or non-governmental organizations, while 17.2% at a university. There were only a small number of respondents (4.9%) working for private/commercial companies in Europe.

The respondents were thus quite representative of European (monolingual) dictionary-making community, considering that in the European survey on dictionary use and culture (Kosem et al., 2019: 96) it was reported that in the majority of the countries participating in the survey monolingual dictionaries are published solely or mainly by public institutions funded by the government.

A total of 58.5% of the respondents were involved in compilation of monolingual dictionaries or databases, either general, specific or dictionaries for learners. Much fewer respondents were involved in compiling bilingual (15.1%), multilingual (13.2%) and dialectal (8.8%) dictionaries or databases.

Sixty-one percent of the respondents have a PhD and the majority have an MA or BA degree in language/linguistics (81.1%). More than one third of respondents (35.8%) have more than 20 years of work experience in the field of lexicography, 24.5% have 10-20 years of work experience and roughly every fifth respondent (20.1%) has 5-10 years of work experience. These responses may be an indication that people who start

working as lexicographers stay in the field for a long time.

More than one third of the respondents (34.6%) have been trained within their own institute, usually by a tutor or senior lexicographer. Roughly every fourth respondent (25.8%) has attended special courses or several courses since starting working in lexicography. Other forms of training attended by the respondents were workshops or summer schools. Only a small number (11.3%) of the respondents reported studying lexicography at university, either as part of an MA course on lexicography or as a special course.

The respondents reported working in teams of different sizes, with relatively similar shares being reported across all team sizes. Overall, the majority of our respondents work in teams consisting of under 10 members, and the predominant team size was 3-6 people (27.4%). More than half of the respondents (56.6%) reported working in a team that consisted only of people from their own institution, and 43.4% reported working together with people outside their institution.

A total of 122 different projects were mentioned by the 158 respondents. Fifty-three of these are permanent projects; these are mainly voluminous monolingual contemporary dictionaries, Wiktionaries, etymological and dialectal dictionaries, as well as a few bilingual dictionaries. Another 18 projects have a duration of 15-20 years; these are also mainly voluminous monolingual contemporary dictionaries, etymological and dialectal dictionaries, as well as bilingual dictionaries.

150 respondents answered the question on dictionary publication format. Out of the 122 reported dictionary projects, 100 (82%) would be published online – 55 of them online only, 45 also in print. For four projects, the respondents reported that dictionaries would also be available as apps. Only 24 projects out of the 122 mentioned in the survey would appear in print only. A reason for publishing in print is tradition; the dictionary is part of a larger project and previous volumes have appeared in print. These results are also in line with what was reported by Kosem et al. (2019) on the status of lexicography (types of dictionaries being compiled and their format) in the 26 countries involved in their study.

The majority of the project databases on which the respondents are working are organized from word to meaning (word-based databases, 87.3%). Databases organized from meaning to word (concept-based, 8.9%) are used mainly in terminological projects. There is also a small number of projects (3.2%) that combine both word-based and concept-based organization of the database.

4.2 Software and tools

Eighty-nine out of 159 respondents answered the question about software and tools. More than half of them reported that they use both a DWS and a CQS in their work.

Altogether 15 DWSs and 22 CQSs were mentioned by the respondents. The tools can be divided into three main categories: commercial, open-source and in-house. General purpose editors, dictionary publishing platforms and App Builders were considered as a separate category.

There are mainly three types of DWS+CWS combinations used by the lexicographers:

1) in-house DWS and commercial CQS (e.g. Ekilex[9] and Sketch Engine[10])

2) commercial DWS and commercial CQS (e.g. IDM[11] and Sketch Engine)

3) in-house DWS and in-house CQS (e.g. LexDF[12] and IMS Open Corpus Workbench).

The first combination listed is also the most common model. Altogether 54.8% of the respondents reported using Sketch Engine as CQS, other CQSs[13] used were, for example, IMS Open Corpus Workbench, CoRes[14], Korp[15], NoSketchEngine, AntConc[16], and COSMAS II[17]. Generally, the lexicographers in our survey reported using one CQS and one DWS, but some respondents use several DWSs (e.g. iLex[18], Lexonomy[19] and TLex[20]) and several CQSs (mostly Sketch Engine in combination with other CQSs such as KonText[21], Lexpan[22] or Korp) at the same time. The following reasons were given for using more than one system: (a) moving from commercial or in-house to open-source; (b) different project needs or needs of lexicographers, e.g. one system is more suitable for retrodigitized dictionaries, another one for born-digital dictionaries; one for word-based, another for concept-based lexicography, etc.

[9] https://ekilex.eki.ee (1 July 2019)

[10] https://www.sketchengine.eu/ (1 July 2019)

[11] http://dps.cw.idm.fr/index.html (1 July 2019)

[12] The product is not publicized, but registered with Inven2, The UiO patent and IPR organization, since 2014.

[13] For the full list see Kallas et al. 2019.

[14] https://korpus.dsl.dk/corest/index.htm (1 July 2019)

[15] https://spraakbanken.gu.se/eng/korp (1 July 2019)

[16] http://www.laurenceanthony.net/software/antconc/ (1 July 2019)

[17] https://www.ids-mannheim.de/cosmas2/ (1 July 2019)

[18] https://issuu.com/jens.erlandsen/docs/ilex_brochure_120dpi (1 July 2019)

[19] https://lexonomy.eu/ (1 July 2019)

[20] https://tshwanedje.com/tshwanelex/ (1 July 2019)

[21] https://kontext.korpus.cz/first_form?corpname=syn2015 (1 July 2019)

[22] http://www1.ids-mannheim.de/lexik/uwv/lexpan.html (1 July 2019)

Relevant for these results are also the findings of the survey for institutions. All but one institution participating in this survey use one or more DWSs and it is still quite common[23] for the institutions to develop an in-house system (five institutions indicated that they use an in-house DWS). It is also not uncommon for the institutions to use more than one DWS because of different project needs. About half of the partner institutions indicated that they did make some adaptations/customizations to an off-the-shelf DWS to make it more suitable for their project(s). The following customizations were mentioned: customization of schemas, DTDs and menus; customization of view options (e.g. for getting an overview of the entry); customization of search and extraction options. All but two institutions use one or more CQSs, often combining a commercial system with an in-house or open-source system.

4.3 Compiling methods and automatic knowledge extraction

All respondents answered the question about compilation methods. As shown in Figure 2, the majority of the respondents reported compiling their dictionaries manually (57.9%).

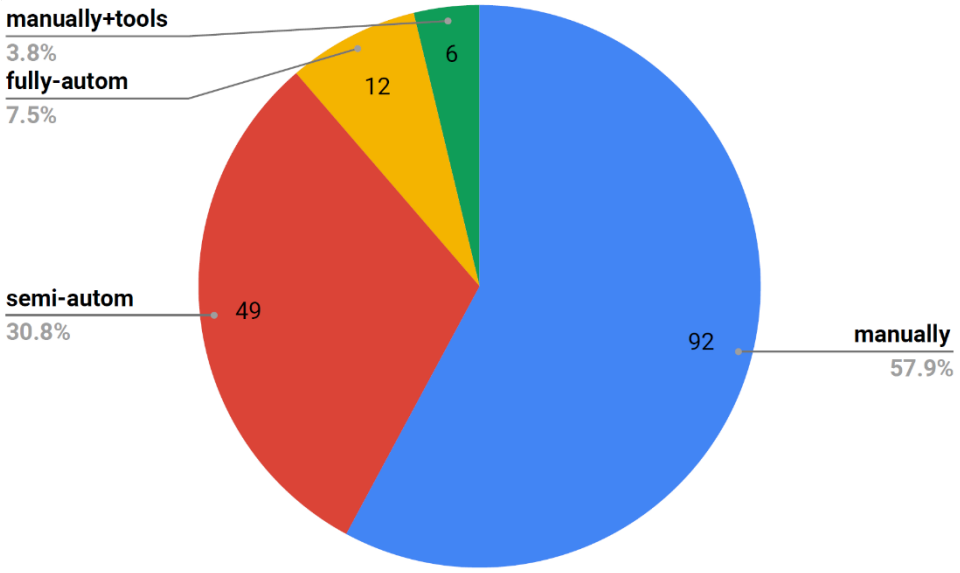

Figure 2: Compiling methods for all projects (N=159)

Based on other answers provided, the respondents perceive the manual method as analysing the data (often by using CQS) and then inserting the information into their DWS or some other tool manually. Nearly one third of the respondents (30.8%) work with semi-automatically collected data , and only 7.5% (12 respondents) are using fully-automatically collected data. It is interesting that the respondents who marked their

[23] As was the case in the COST ENeL survey (Tiberius & Krek, 2014).

project to be born-digital[24] (N=65) mentioned using different compiling methods: mainly semi-automatic (43.1%) and manual (!) (33.8%). The most common types of data for which the respondents reported using automatic extraction methods include headword list (20.8%), collocations (12.7%) and frequency information (11.3%). Automatic extraction of multi-word expressions (8%), dictionary examples (7.5%) and form variants (6.1%) is fairly common, too. Less than 5% of the respondents, respectively, reported using automatic extraction for patterns (4.7%), neologisms (3.8%), lexical-semantic relations (3.8 %), domain information (4.4%), multilingual data from parallel/comparable corpora (3.8%), definitions (3.3%) and audio data from speech corpora (2.4%).

4.4 Various aspects of the lexicographer's job

In this part we have chosen four different aspects of the lexicographer's job that feed back into the ELEXIS projects in terms of training, IT support, user involvement and tools for retrodigitization. Namely, all these aspects are important in state-of-the-art lexicography as the job of a lexicographer has changed – most lexicographers do not just edit dictionary entries anymore, but are also involved in various other aspects of dictionary-making.

4.4.1 Involvement of lexicographers in the online publication process and user research

Lexicographers were asked to specify what kind of work they are doing when they are involved in online publication or user research. It was an open-ended question, but three options were proposed: 1. evaluating the user interface and providing new ideas; 2. creating add-on materials (e.g. blogs, slideshows, videos, quizzes, word games); 3. communicating with IT persons / user experience designer (UX) / interface designer (IX).

Just over a quarter (27%) of 63 respondents answered that they are not involved in online publication, while the 33.9% who were involved in online publication dealt with user interface evaluation, and communication with IT specialists, including user experience designers and interface designers. In addition to user interface evaluation

[24] Furthermore, terms such as "born-digital" and "IT support" seem to have been interpreted in different ways by different respondents, even although a definition of "born-digital" was provided. For example, the share of respondents who answered the question whether they work on born-digital dictionaries affirmatively was unusually high, especially considering the information they provided for related questions about the types of projects, compilation methods and the format of publication, which suggest a different interpretation of the term "born-digital". This experience shows not only that all terms should be defined in future surveys, but also that there is a need for a discussion of the term in the lexicographic community, something that the ELEXIS project should also pay attention to.

and communication with IT specialists, 16.9% of the respondents were involved in the production of add-on materials. Just 11.9% are involved only in user interface evaluation and 8.5% only in IT communication. Other tasks mentioned include project management, updating user guides, organizing and testing new editions (or updates of existing editions), working on promotional activities (e.g. media interviews, presentations, Word of the Day), analysis of user feedback, answering user questions, etc.

The respondents were also asked if they are involved in user research for their projects, and if so what kind of user research they do. The options proposed were, for example, analysing user logs or interviewing end users. Just under two thirds (62.5%, or 55 out of 120 respondents) revealed that they are not involved in user research. A total of 59% of those lexicographers who do user research conduct analyses of user logs, 33.2% also conduct interviews with end users (mostly before and during the conceptual phase of the dictionary). Other tasks mentioned include the analysis of data from language-related advisory services and Google Analytics, the analysis of user feedback, mostly proposals and corrections (the feedback is gathered through mail or online feedback forms), conceiving and supervising user studies carried out by others, and informal consultation.

4.4.2 IT support

As expected, IT support is an important part of lexicographer's job. Over 80% of the respondents answered this question and reported to have either basic (43.9%) or good (37.8%) IT support. We did not look into the dynamics between lexicographers and IT staff in more detail in this survey, but it definitely deserves more attention, particularly the way in which IT staff are perceived by lexicographers, and whether there are differences in the way the lexicographers perceive IT staff or computational linguists or NLP experts. IT tasks are also the only tasks that seem to be outsourced in dictionary projects, ranging from designing the online interface of the dictionary to developing and/or offering support in the use of DWS or CQS. Trustworthy experts, efficiency and another view of the data and content (which might help to identify some lexicographic problems) were mentioned as positive experiences. The cost (too expensive, lack of (regular) funding), more additional work (to teach and explain lexicographic details), delays and communication problems were mentioned as negative experiences when outsourcing.

4.4.3 Crowdsourcing, gamification and data enrichment

The results of the survey show that crowdsourcing and gamification are not yet common practices in the lexicographic projects that the respondents are involved in. Nonetheless, the wish for tools for crowdsourcing was put down by several respondents in the survey. These results are not that surprising, as crowdsourcing has become a hot topic in

lexicography only in the last five years, so it is understandable that many projects (and lexicographers) are still cautious about using the wisdom of the crowd.

Of particular interest are the results of the question related to data enrichment (i.e. adding additional linguistic and non-linguistic information to the data, such as normalizing values, geo-locating, expanding content, etc) which not only concerns retrodigitized dictionaries, but also born-digital dictionaries which can be enriched with various types of information. Different forms of data enrichment were mentioned in the context of retrodigitization by the respondents, e.g. text normalization, expanding abbreviations, adding grammatical information as well as adding internal and external links. Relatedly, the survey for institutions showed that data enrichment is not yet very common in current lexicographic projects within the ELEXIS consortium. Only two institutions indicated that they include images and/or videos in their dictionaries.

4.4.4 Retrodigitization

As retrodigitization of older/printed dictionaries (i.e. the process of converting a dictionary published in paper into a digital, computer-readable format, which involves not only scanning and OCRing but also data encoding and enrichment) is an emerging trend in modern e-lexicography, we asked the respondents about their involvement in different phases of the retrodigitization process. The aim was to get an overview of the software used in this process and to collect lexicographers' opinions on which dictionaries should be retrodigitized. The number of the respondents, 16, that answered these questions was rather low. This may be due to the fact that some parts of the retrodigitizing activities (image and text capturing) are not directly related to the lexicographic work. If we look at the individual phases of retrodigitization, we see that the 16 respondents reported to be mainly involved in the activities which require lexicographic competence, such as data encoding (15 responses) and data enrichment (13 responses).

5. Discussion

5.1 Lexicographer's training and job description

The information on the experience and training of the respondents of the survey points to another potential issue in lexicography. Although the respondents reported having quite a lot of experience in lexicography, they all had to be trained by their employer; very few of them actually had a formal education in lexicography. Consequently, dictionary-makers have to be prepared for extra costs related to training of their staff, and need to plan projects accordingly.

This situation makes degree programmes such as EMLex (European Master in

Lexicography)[25] very important for the training of young generations of lexicographers, and the development of the field in general. At the same time, it is essential that lexicographers are provided with different types of quality training materials, something that ELEXIS is also dedicated to provide as part of Work Package 5.

Education and training of lexicographers will need to become more and more interdisciplinary, as the findings of the survey indicate that a lexicographer's job is far from being monotonous. Modern lexicographers need to possess much more than just linguistic skills; other skills in their repertoire need to include project management, communication with computational staff, promotional activities, responding to user questions and feedback, etc.

It is noteworthy that most lexicographers are not involved in the final dictionary publication (of an online dictionary) or user research. The former finding is to be expected, as normally this job is left to web/interface designers; however, one cannot help wonder whether dictionaries are really better for it. On the other hand, the lack of at least some involvement into user research is worrying, especially considering the current lack of user research in most European countries (and around the world) (Kosem et al., 2019: 96). Knowing the users is important; as Atkins and Rundell (2008: 5) rightly point out, "the content and design of every aspect of a dictionary must, centrally, take account of who the users will be and what they will use the dictionary for". Part of the solution might be in conducting regular European- or world-wide surveys, such as Kosem et al. (2019) and Müller-Spitzer (2014), as this brings lexicographers together and also promotes the discipline (and dictionaries) among the general public.

5.2 Existing tools and lexicographers' wishes for the future

As shown in Sections 4.2 and 4.3, the lexicographers use a wide variety of DWSs and CQSs, in different combinations. Moreover, they often use more than one DWS or CQS, mostly because of the needs of specific dictionary projects. The finding that an in-house solution is the predominant form of DWS used is in line with the findings of the ENeL survey (Krek et al. 2014). It thus seems that existing off-the-shelf DWS often still do not meet (all) the needs of lexicographic projects.

As the development of new open-source tools is an important part of the ELEXIS project, it was also important to learn about the respondents' wishes regarding DWS and CQS, in other words what would be their ideal tool. The majority of the respondents mentioned that their ideal DWS should be free, online, open-source, browser independent, fast, intuitive, and easy to maintain. This supports the view of ELEXIS and reaffirms our aims to develop online open-source tools such as Lexonomy.

[25] https://www.emlex.phil.fau.eu/ (1 July 2019)

Other features, such as supporting real-time collaborative input, real-time saving, localization, customizability both in terms of functionalities and interface, online publishing of the results, and proper documentation (i.e. it should not be a black-box system) were also listed. While many existing DWSs already have most of these features listed by the respondents, it seems that all of them are mandatory as far as lexicographers are concerned.

The respondents also believe it is important that their DWS is interoperable with other resources, operating systems and tools. Thus, API and script support is expected. This was mentioned both in connection with the possibility of automatic pre-compilation of entries and the possibility to integrate lexicographic information automatically from CQS into DWS. Similar findings were observed in the survey for institutions, where most partner institutions felt that the integration of DWS and CQS would be beneficial, especially for the linking, selection and retrieval of examples, collocations, etc. Again, this is something that the ELEXIS project is working on addressing as, at the time of writing this paper, the beta version of the Sketch Engine pull feature in Lexonomy was already available. The feature enables quick search and import of examples, collocations, synonyms, and even definition candidates (for some corpora) from Sketch Engine into Lexonomy.

In terms of CQS, the answers from the institutional survey are also important to mention here, as the respondents listed some features that they missed in existing CQSs, such as sense clustering (clustering concordances against senses)[26], implementation of syntactic and semantic annotation, detection of neologisms, automatic acquisition of translation equivalents, diachronic analysis, etc. The topicality of these features is also evidenced by the fact that the ELEXIS project contains various activities focussed on these.

Relatedly, several respondents also pointed out the need for better tools for retrodigitization. Such tools include automatic processes where the quality of output highly influences the amount of manual labour needed to prepare the digital version of the dictionary, for whatever purpose it is then used.

5.3 Current trends and looking ahead

It can be said that automatic knowledge extraction in lexicography is definitely on the increase, and the findings of this survey are very much similar to the findings of the ENeL survey in 2014-2015 (Krek et al., 2014). Also, headword lists, frequency information and multiword expressions (collocations in particular) are still the most commonly extracted types of information. Less common automatic extraction of

[26] The respondents might have been influenced by the formulation of the question, as this was one of the suggestions listed to help them understand the question.

information that is more semantically-based, such as senses, definitions, lexical relations, etc., can be attributed to the fact that lexicographers do not seem to think that existing tools already perform these tasks satisfactorily enough; this is evidenced in the respondents' answers to the question on their needs in the next 10-15 years, where the most mentioned topic was the need for better tools for extraction and automatic processing of data from corpora.

Moreover, lexicographers seem to be well aware of the potential of the Semantic Web, Linked Open Data, and Artificial Intelligence for lexicographic purposes.

One of the things that the respondents reported had improved was the interaction between the users and the dictionary, since users can now directly contact lexicographers online about words they are looking for, technical issues, etc. At the same time, the respondents called for more and better tools to analyse user behaviour. Considering the poor status of user research in many countries (Kosem et al., 2019) and lack of lexicographer involvement in user research (reported in the survey presented in this paper), such tools and probably training to help facilitate research into dictionary use should definitely be provided.

Two of the emerging trends in lexicography are crowdsourcing and gamification; however, at the moment their use is largely limited to user feedback (e.g. mistakes in entries or suggestions for new words). The use of crowdsourcing during dictionary compilation is used by only a few lexicographic institutions and projects, for example the Thesaurus of Modern Slovene (Arhar et al., 2018), the Collocations Dictionary of Modern Slovene (Kosem et al., 2018), the Estonian project for the dictionary of associations (Vainik, 2018) and the Taalradar project[27] at the Dutch Language Institute, but in those cases it has proven to be very effective. Still, progress from the situation reported in the ENeL survey in 2014 (Krek et al., 2014) can definitely be observed. But overall, it seems that lexicographers are still searching for the best ways of including these methodologies in dictionary compilation. Potential issues could be the lack of suitable case studies, and the lack of relevant features in existing DWS or the lack of tools supporting these methods. This need was also reported by several respondents in this survey.

Among other relevant wishes expressed by the respondents that deserve to be mentioned are the need for a common standard for the development of lexicographic resources, the need for a central repositorium, and the need for tools for harmonization of dictionary formats. The respondents expect a significant change in relation to lexicographic data modelling and publishing policy. The turn towards unified data is expected, with respondents mentioning that publishers will produce a single resource containing all the data that the publisher has about the language, including data traditionally not considered part of a dictionary. Considering that providing solutions

[27] https://taalradar.ivdnt.org/

to these issues is also part of the ELEXIS agenda, it is good to see that the lexicographers are aware of them.

It is also important to note some of the concerns that were expressed by the respondents. These were often connected to the quality and reliability of lexicographic data in state-of-the-art lexicography, information overload, rapid technology development, and the potentially reduced value of lexicographic skills in digitally oriented projects. Several respondents were concerned about the overestimated value of the presentational component of dictionaries, especially in relation to presentation on smartphones, which may result in neglecting the aspect of the quality and reliability of lexicographic data. Last but not least, a few respondents noted the low status of lexicography in their countries. This echoes events such as the recent discontinuation of important national dictionary projects (e.g. Great Dictionary of Polish; Żmigrodzki, 2018), reports on the absence of teaching dictionary use in schools (Kosem et al., 2019), and acceptance of documents such as the Resolution at EURALEX 2016 Congress, promoting the importance of lexicography.

6. Conclusion

The survey conducted as part of the ELEXIS project has provided useful insights into existing practices and needs of lexicographers around Europe. The survey successfully complements the surveys conducted during the ENeL COST Action, especially in terms of raising awareness of issues such as lexicographer education and training, lexicographers' needs connected with tools, and the latest lexicographic trends. It also points to the importance of regular updating of information about the lexicographic practices, methods, tools and formats used in institutions across Europe and the world.

We intend to collect more data on lexicographic practices in the coming years, e.g. by including the ELEXIS observer institutions and their lexicographers in a follow-up survey. In this way, we intend to devise some form of a lexicographic practice map of Europe so that similarities and differences between practices at different institutions in different countries can be easily analysed. This would facilitate institutional collaboration and the search for common solutions. Finally, all the results have already informed and will continue to inform the preparation of the deliverables of the ELEXIS project, such as tools, resources and training materials that will be produced in the next three years.

7. Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

8. References

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. & Robnik Šikonja, M. (2018). Thesaurus of modern Slovene: by the community for the community. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress, 17-21 July 2018, Ljubljana. Ljubljana: Ljubljana University Press, Faculty of Arts*, pp. 401-410. http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2991-1-10-20180820.pdf.

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Hartmann, R.R.K. (ed.). (2003). *Lexicography: Dictionaries, compilers, critics, and users.* London: Routledge.

Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). ELEXIS deliverable 1.1 Lexicographic Practices in Europe: A Survey of User Needs. https://elex.is/wpcontent/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (22 February 2019).

Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In. H.R. Gouws, U. Heid, W. Schweickard & H.E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin–Boston: de Gruyter, pp. 517–524.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress, 17-21 July 2018, Ljubljana. Ljubljana: Ljubljana University Press, Faculty of Arts*, pp. 989-997. http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2939-1-10-20180820.pdf.

Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M. & Wolfer, S. et al. (2019). The image of the monolingual dictionary across Europe. Results of the European survey of dictionary use and culture. *International Journal of Lexicography*, 32(1), pp. 92-114. https://doi.org/10.1093/ijl/ecy022 (30 May 2019).

Krek, S., Abel, A. & Tiberius, C. (2014). *Dictionary Writing Systems & Corpus Query Systems. Survey – WG3 ENeL.* http://www.elexicography.eu/wp-content/uploads/2015/04/ENeL_WG3_Vienna_DWS_CQS_final_web.pdf (30 May 2019).

Müller-Spitzer, C. (ed.) (2014). *Using Online Dictionaries.* Berlin, Boston: De Gruyter.

Tiberius, C. & Krek, S. (2014). *Workflow of Corpus-Based Lexicography.* Deliverable COST-ENeL-WG3 meeting, July 2014, Bolzano/Bozen. http://www.elexicography.eu/wp-content/uploads/2015/04/LexicographicalWorkflow_DeliverableWG3BolzanoMe

eting2014.pdf (30 May 2019).

Tiberius, C., Heylen, K. & Krek, S. (2015). *Automatic Knowledge Acquisition for Lexicography. Survey – WG3 ENeL.* http://www.elexicography.eu/wp-content/uploads/2015/10/ENeL_WG3_Survey-AKA4Lexicography-TiberiusHeylenKrek.pptx (30 May 2019).

Vainik, E. (2018). Compiling the Dictionary of Word Associations in Estonian: from scratch to the database. *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian Papers in Applied Linguistics* 14, pp. 229—245. http://dx.doi.org/10.5128/ERYa14.14 (30 May 2019).

Żmigrodzki. P. (2018). Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress, 17-21 July 2018, Ljubljana. Ljubljana: Ljubljana University Press, Faculty of Arts*, pp. 209-219. http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2973-1-10-20180820.pdf (30 May 2019).