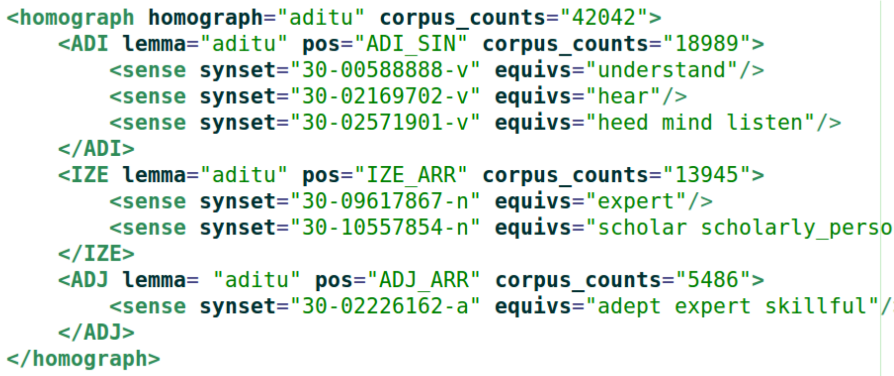
Table of contents

Linking Basque Lexical Resources

Motivation Letter for LexMC by David Lindemann

I am currently working in Hildesheim as postdoc researcher-lecturer at the Computational Linguistics chair of Prof. Ulrich Heid. I am involved in defining theoretical and methodological preliminaries for a project on **Linking Basque Lexical Resources**. After publishing a preliminary study, it will be my intention to apply for funding, in order to carry out a research project of several years, starting eventually in the beginning of 2019, at the public UPV/EHU University of the Basque Country, where I have good relations both at the Faculty of Arts (Basque Philology, Basque lexicography), and the Faculty of Computer Science (Basque Computational Linguistics, NLP resources development). The project may be outlined as follows.



```
<homograph homograph="aditu" corpus_counts="42042">
    <ADI lemma="aditu" pos="ADI_SIN" corpus_counts="18989">
        <sense synset="30-00588888-v" equivs="understand"/>
        <sense synset="30-02169702-v" equivs="hear"/>
        <sense synset="30-02571901-v" equivs="heed mind listen"/>
    </ADI>
    <IZE lemma="aditu" pos="IZE_ARR" corpus_counts="13945">
        <sense synset="30-09617867-n" equivs="expert"/>
        <sense synset="30-10557854-n" equivs="scholar scholarly_person bookman"/>
    </IZE>
    <ADJ lemma= "aditu" pos="ADJ_ARR" corpus_counts="5486">
        <sense synset="30-02226162-a" equivs="adept expert skillful"/>
    </ADJ>
</homograph>
```

Fig. 1: Merged Basque resources: Abbreviated sample.
POS tagset: ADI (verb), IZE (noun), ADJ (adjective)

Starting point: As explained in our Euralex 2016 paper (Lindemann and San Vicente, 2016), we may merge existing Basque lexical resources in order to obtain a Basque dictionary draft that includes lemma signs („homographs“), lempos-entities, word senses, and lexicalisations for these word senses in Basque, English and via aligned wordnets in several other languages, as well as corpus frequency information at lemma sign and lempos level (references in table 1). A sample entry of this dictionary draft may be represented in XML as shown in Fig. 1 above. With the described dictionary draft on hand, our proposal for a workflow includes detection of gaps in the draft, i.e., of cases where a EDBL lempos-entity finds no correspondance in EusWN or vice versa, of cases where a frequent lemma-sign corresponds to no known lempos-entity or EusWN lexicalisation, etc., and lexicographic editing work.

EusLemStd (Lindemann and San Vicente, 2015)	Corpus based frequency lemma list	53,000 homographs (lemma signs)
EDBL (Aldezabal *et al.*, 2001)	Basque lempos repository (pos tagger, spell checker lexicon)	84,000 lempos entities
EusWN (Pociello, Agirre and Aldezabal, 2011)	Basque WordNet	50.000 lexical items
ETC (Sarasola, Landa and Salaburu, 2013)	Basque reference corpus (hand selected)	200M tokens
Elh200 (Leturia, 2014)	Basque WaC	200M tokens

Table 1: Some existing resources for Basque

A. First idea to develop: The content of two major Basque reference dictionaries that follow actual Standard Basque lemmatisation and orthography rules are or will be represented as XML: (a) the very large OEH *General Basque Dictionary* (Mitxelena and Sarasola, 1988), and (b) the Basque monolingual *Euskal Hiztegia* (Sarasola, 1996). Samples of both are available. My intention is to propose (i) a data model that allows a merging (or linking) of the lemma signs, lempos-entities and word senses present in both resources to the basic model shown in Fig. 1, so that the microstructural content attached to these entities can be represented as part of the main resource, as shown in Fig. 2, either attached to a homograph, a lempos, or a word sense element; and (ii), a workflow that addresses related issues, like resolving homonymy vs. polysemy, and, for polysemous lempos-entities, identifying overlapping word senses accross resources as candidates for merging, etc.

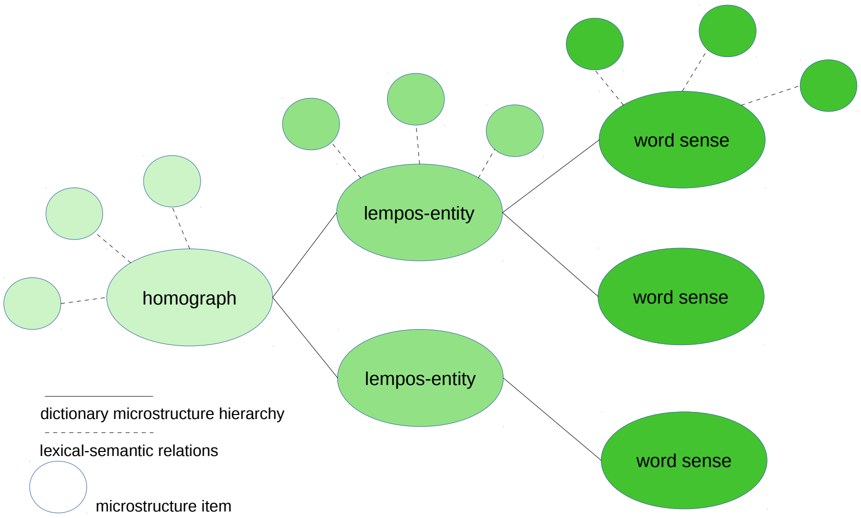


Fig. 2: Microstructural elements and relations

B. Second idea to develop: As explained in our eLex 2017 contribution (Lindemann and Kliche, 2017), it is our intention to propose a data model and a lexicographic workflow that allows (1) a drafting from scratch of bilingual dictionaries with Basque using wordnets aligned at synset (concept) level, (2) a manual editing of this draft, and (3), thought as reciprocal process, an update of EusWN's sense (concept) repertoire „through the back door“, with the goal of ultimaltely having (4) an bootstrapping loop spinning between EusWN and ongoing bilingual dictionary projects.

On the LexMC workshop, first and foremost I would like to put forward (**A**), by working on the actual dictionary samples, discussing the placement of macro- and microstructural elements in a merged data model, but also discuss issues that will arise related to (**A**) and (**B**), i.e., for example, the presentation of word senses in monolingual vs. in multilingual dictionaries (asymmetries as found for the German noun *Ufer*, that counts one single sense in the monolingual DUDEN, while in a German-English dictionary it counts two, depending it is a *(sea) shore* or a *(river) bank*) or in concept-based wordnet-like resources vs. in lemma-based dictionaries.

Although I feel already quite familiar with some of the workshop contents, I still constantly feel a need for deeper and systematic formation in technical aspects, related programming skills, etc., and I am pretty sure that all teaching topics will contain new information for me I can immediately make use of. As for the format of the workshop, I am very sure that I will learn a lot from working together with the trainers and the other participants.

References:

Aldezabal, I. *et al.* (2001) 'EDBL: a General Lexical Basis for the Automatic Processing of Basque', in *Proceedings of the IRCS Workshop on linguistic databases*. Philadelphia. Available at: http://artxiker.ccsd.cnrs.fr/docs/00/08/11/54/PDF/2001-IRCS.pdf

Leturia, I. (2014) *The Web as a Corpus of Basque*. PhD Thesis. UPV-EHU Lengoaia eta Sistema Informatikoak Saila.

Lindemann, D. and Kliche, F. (2017) 'Bilingual Dictionary Drafting: Bootstrapping WordNet and BabelNet', in *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017*. *eLex 2017*, Leiden: Lexical Computing, pp. 23–42. Available at: http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202016/euralex_2016_099_p898.pdf

Lindemann, D. and San Vicente, I. (2015) 'Building Corpus-based Frequency Lemma Lists', *Procedia - Social and Behavioral Sciences*. (Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015).), 198, pp. 266–277. Available at: http://www.sciencedirect.com/science/article/pii/S1877042815044468

Lindemann, D. and San Vicente, I. (2016) 'Bilingual Dictionary Drafting: Connecting Basque word senses to multilingual equivalents', in *Proceedings of EURALEX 2016*. *XVII International Euralex Conference*, Tbilisi: Tbilisi State University, pp. 898–905. Available at: http://euralex2016.tsu.ge/publication2016.pdf

Mitxelena, K. and Sarasola, I. (1988) *Diccionario general vasco - Orotariko euskal hiztegia*. Euskaltzaindia; Editorial Desclée de Brouwer. Available at: http://www.euskaltzaindia.net/oeh XML conversion planned / in course.

Pociello, E., Agirre, E. and Aldezabal, I. (2011) 'Methodology and construction of the Basque WordNet', *Language Resources and Evaluation*, 45(2), pp. 121–142. doi: 10.1007/s10579-010-9131-y.

Sarasola, I. (1996) *Euskal Hiztegia*. Donostia: Kutxa Gizarte-eta Kultur Fundazioa. XML version: Arriola, J. *et al.* (2003) 'Semiautomatic conversion of the Euskal Hiztegia Basque Dictionary to a queryable electronic form', *T.A.L. journal*, (44:2), pp. 107–124.

Sarasola, I., Landa, J. and Salaburu, P. (2013) 'Egungo Testuen Corpusa'. UPV-EHU. Available at: http://www.ehu.es/etc/

David Lindemann
University of Hildesheim
IWiSt Institute for Information Science and Natural Language Processing
Universitätsplatz 1, D-31141 Hildesheim
[Bühler-Campus, Lübecker Str. 3, LN131]
Tel. +49(0)512188330-336
web https://www.uni-hildesheim.de/~linde002/
email david.lindemann@uni-hildesheim.de

Application for the Master Class

I am first year PhD student in language technology in the University of Helsinki. My work is strongly related to lexicography, but from a computational point of view. In my research, I work with existing XML lexica for small Uralic languages. My goal is to bring these dictionaries to a crowdsourced MediaWiki environment[1] so that people with no technical or linguistic background can profit from these dictionaries by just viewing them or by editing them.

Another goal of my research is to expand these lexical resources automatically by combining translations from different lexica of different languages and also by automatically extracting interesting lexicographical information from corpora. An automated corpus based approach can be used for example to find cognates automatically or to expand the semantic knowledge of the lexica.

I acknowledge that my view point to lexicography is computational. This is why I think this master class can be useful for me to learn more about the linguistic aspect of lexicographical research. I am also eager to learn whether the problems we are facing with our XML lexica are common in the field and whether the automated methods I am developing in my research can be useful for other scholars in the field.

I am especially eager to learn about finding examples to be used in a dictionary. This is also a question we have been considering worth an automatic approach. This it would be useful to know how these examples are usually picked by experts and how this could be turned into a heuristic for a computational system that can extract quality examples from a corpus.

The dataset I would like to look more closely into during the master class is the XML lexica for Skolt Sami[2] in the Giellatekno infrastructure. This lexical dataset consists of multiple XML files each

[1] Sanat dictionary platform in the Finnish CSC infrastructure, see https://sanat.csc.fi/wiki/Sms:sokk
[2] https://victorio.uit.no/langtech/trunk/langs/sms/src/morphology/stems/

of which corresponds to one part-of-speech, i.e. there is an XML for adjectives, another one for adverbs and so on.

The XMLs consist of entries. Each entry refers to a lexeme and contains information about the lexeme, at least its lemma and a translation, but usually much more information is provided. The additional information can, for example, be etymology, continuation lexicon for morphological analyser, semantic tags, audio samples etc. An interesting thing in the translations is that they are always divided into so called meaning groups. A meaning group contains translations that refer to a single semantic concept. In the case of polysemous words, each meaning has its own meaning group.

Currently, I am working with an automated approach to merge meaning groups from lexica of different languages in such a fashion that we could provide more translations for each minority language and also provide translations in completely new languages. This is a work in progress that I could continue in the context of the master class.

The XMLs are freely available and all my research in expanding them automatically will be made available in the very same open-access Giellatekno repository the original versions currently are located at.

Russian manuscript lexicons have been the object of my investigation for quite a long time: I was studying them while writing my PhD thesis “David Zamaray’s lexicon: its lexicographical, typological and structural peculiarities” (to be defended early in 2018) and working at the individual research project “Russian manuscript dictionaries as a lexicographical genre of Muscovy: traditions and innovations”, supported by the Russian Humanitarian Scientific Fund (RHSF) in 2013-2015. In the process of my research I encountered to the problem of getting access to the manuscripts I needed, as they were stored in various archives all around Russia: in Moscow, St. Petersburg, Novosibirsk, Tomsk, Vladimir, and Yaroslavl. Visiting archives is time consuming, and facsimiles of the manuscripts are quite expensive. Still, financial support of RHSF made it possible to me to work in the main Russian archives, copy some of the lexicons in a plain text format and obtain some facsimiles. I am convinced that the lexicons should be represented online as a database supplied by some additional information about each manuscript and words explained there and supported by a search engine. I suppose that the most suitable format here is TEI markup as word entries do not always have strict structure and cannot be represented in a relational database.

As the material for the master class, I would propose a lexicon from the Russian National Library in St. Petersburg (Solovki collection, № 302/322), representing an early stage of the Russian lexicography. It was created at the beginning of the 17th century as an addition to the “The Ladder of Divine Ascent” of John Climacus and consists of five smaller lexicons of about 1,000 word entries in total. The title of the lexicon – “An Explanation of the words difficult for understanding in texts, as these words are translated by the first interpreters some in Church Slavonic, some in Serbian, others in Bulgarian, some in Greek and other languages and which were not translated into Russian” – reflects the purpose of the compilation: to explain foreign words which were just transliterated during the process of translation and could be found by readers in the Old Russian texts. Most such words have Greek origin (*Агиосъ* < Gr. *ἅγιος*, *Кюриосъ* < Gr. *κύριος*), some – of Latin (*Генезисъ* < Lat. *Genesis*, *креаторъ* < Lat. *creator*), Hebrew (*Аданаи* < אדני, *Бресшивъ* < בְּרֵאשִׁית), Ruthenian (*жолнеръ*, *крыжсъ*) or Church Slavonic (*ветия*, *скуделникъ*). Some words have information about their origin (language marks) and the title of the literary sources where the words came from (books of the Holy Scripture, works of the Holy Fathers, or biographies of the Orthodox saints). The explanatory part consists mostly of the Russian equivalent or description of the notion[1].

TEI markup of the word entries could be as follows:

[1] More detailed information about the structure of word entries and the way they changed can be found in the article «The Word Entry Structure of the Russian Manuscript Lexicons: Evolution through the Centuries» (https://www.academia.edu/31388772/The_Word_Entry_Structure_of_the_Russian_Manuscript_Lexicons_Evolution_through_the_Centuries_Proceedings_of_the_XVII_EURALEX_International_congress._Lexicography_and_Linguistic_Diversity._Tbilisi_2016._P._584-590)

```
<entryFree>
        <form norm="патриарси" corresp="#495">Патрї́арси <note place="margin"><bibl
n="Acts17"><title key="Acts">дѣѧн</title> <num
value="17">з҃і</num></bibl></note></form>
        <sense> (т) <def>ѿц҃ем нача́лницы.</def></sense>
</entryFree>
<entryFree>
        <form norm="патриархъ" corresp="#496">Патриархъ.</form>
        <sense><def>ѿц҃емъ нача́ло.</def></sense>
</entryFree>
<entryFree>
        <form norm="рексъ" corresp="#607">Реѯ.</form>
        <sense><def>ко<lb/>ро́ль.</def></sense>
</entryFree>
```

Besides, there supposes to be another document with additional data about the headword, in which its normalized form, origin, foreign etymon, grammatical information and topic group will be given:

```
<entryFree>
        <form xml:id="495" norm="патриархъ">патриарси</form>
        <gramGrp><pos>noun</pos><gen>m</gen><number>pl</number></gramGrp>
        <lang>grc</lang>
        <etym>πατριάρχης</etym>
        <sense>патриарх</sense>
        <lbl>Church Hierarchy</lbl>
</entryFree>
<entryFree>
        <form xml:id="496" norm="патриархъ">патриархъ</form>
        <gramGrp><pos>noun</pos><gen>m</gen><number>sg</number></gramGrp>
        <lang>grc</lang>
        <etym>πατριάρχης</etym>
        <sense>патриарх</sense>
        <lbl>Church Hierarchy</lbl>
</entryFree>
<entryFree>
        <form xml:id="607" norm="рексъ">рексъ</form>
        <gramGrp><pos>noun</pos><gen>m</gen><number>sg</number></gramGrp>
        <lang>lat</lang>
        <etym>rex</etym>
        <sense>король</sense>
        <lbl>Rank</lbl>
</entryFree>
```

The second document with additional information will give an opportunity to find, for example, all word entries explaining terms of Church hierarchy or having the same word in different forms (in singular and in plural), all headwords in plural form, all headwords of Greek or Latin origin, etc.

At the same time, a lot of other technical stuff (such as XSLT, XQuery, eXist-db and so on) is required in order to represent the lexicon online with the possibility to make the above-listed inquiries. I hope that the master class will give me the skills how to use all these (or maybe other) technical instruments to overcome the gap between a text with the TEI mark up and the multifunctional infrastructure available online.

The lexicon “An Explanation of the words…” was not published before, and no copies of it are known. It is just one of more than 150 lexicographical compilations of the type known by researchers, but its digital representation might become the first step in creating a database of the Russian manuscript lexicons.

I would like to submit my application for the LexMC: Lexical Data Master Class. My name is Ellert Thor Johannsson and I am a dictionary editor at the *Ordbog over det norrøne prosasprog/A Dictionary of Old Norse Prose* (ONP), which is a dictionary project hosted at the University of Copenhagen and part of the Arnamagnæan Institute. My academic background is in Historical Linguistics, Germanic and Icelandic. I received my undergraduate degree from the University of Iceland in 1996, a MA degree in General Linguistics from Cornell University in 2003 and PhD from the same institution in 2009.

I have held my current editorial position at ONP since 2012, but have been involved in the ONP project since 2006. During this time, ONP has evolved from a print to a digital publication and established itself as an online lexicographic resource. We the editors are continually striving to improve the online edition, by adding material and features, making it relevant for a wider community of users.

I would be attending this course together with my colleague Simonetta Battista. In the past four years we have been very active in promoting ONP within the larger lexicographic community as well as in the academic context of Old Norse studies. We have participated in various international conferences, such as E-Lex, Euralex, ICHLL, NFL, attended seminars and courses, such as Lexicom (Sketch Engine) workshop and published numerous articles on various aspects of the ONP dictionary.

Why we want to attend this course

What we want to focus on during this Masterclass is how to implement some of the ideas we have for improving our dictionary. ONP is characterized by a large collection of citations and very detailed information about the material. Currently we work with the dictionary material in a database environment where the data are arranged in a variety of different interlinked tables and accessed through different sql-queries. Although we are used to working with our data in this way we are very much aware of how xml is being used in lexicographical work and serves as the basis of many popular editing systems. Therefore we are eager to increase our knowledge about the possibilities of its use.

Besides our general interest in learning more about the xml-based approach to dictionary structuring and editing we have two main interests. Firstly we would like to learn about ways to integrate xml-marked up data into our dictionary system. The main reason is that in recent years the amount of available electronic scholarly editions of Old Norse texts has increased tremendously and many of them are encoded in xml. Furthermore a good number of these editions include lemmatization of the text which makes them optimal to use in a dictionary context. Some texts are even freely available online, such as the texts found in the Medieval Norse Text Archive (menota.org). We would like to be able to extract such lemma information from the xml-files of the editions and arrange them in a dictionary entry structure, similar to the one we use for the ONP. We would also like to create a link between such an xml-created structure and the lemma list from our dictionary. Integrating such resources into ONP Online would provide additional citations to structured dictionary entries, as well as a wider variety of searchable examples of word use.

Secondly we would like to gain better insight into some of our dictionary data. The majority of the citations found in our database are a result of selective excerpting of scholarly text editions. We do not know how representative they are of the vocabulary of particular texts or periods, as the excerpting was only governed by the individual editors' judgment. We would like to figure out how

the vocabulary of Old Norse is represented in the ONP database and which words have been selected in the early excerption process. In order to facilitate this research we would like to analyse freely available xml-files containing lemmatized non-scholarly editions of some essential saga-texts and compare them to the corresponding citation material found in the ONP database.
The material we are bringing with us:

- Two fully lemmatized Old Norse texts in xml-form: *Strengleikar* (38453 words), *Barlaams saga* (76411 words), which are freely available in the Medieval Nordic Text Archive (menota.org).
- Complete ONP wordlist in table (and potentially xml) format
- Two saga texts, *Grettis saga* and *Fóstbræðra saga* in xml, normalized to Modern Icelandic and lemmatised using automated linguistic analysis (92.7% accuracy).
- Subset of citations from ONP in table format (and potentially xml) from those two sagas.

We, the editors of ONP, believe that by participating in this course we would gain important insight into the xml-based tools available for lexicographic work and would receive valuable training working with creating, managing and using digital lexical data. We are sure this experience would benefit the ONP dictionary in the long run and also enable closer work with the Arnamagnæan Institute's current focus on digital scholarship, as is evident by such xml-based projects as *Script and Text in Time and Space* and the *Stories for all time: The Icelandic Fornaldarsagas*, spearheaded by Prof. Matthew Driscoll.

Background of the ONP project

In what follows is a brief historical overview of the project to show where we are coming from. This overview puts the project in context and helps explain the nature of the ONP dictionary data, the development of the dictionary through the years and the current state of affairs.
The project was established in 1939 as a new, historical, scholarly dictionary which should provide an original lexical description of Old Norse with a large number of examples of word use and various kinds of supplementary secondary information.
The source material consists of Icelandic and Norwegian medieval texts, from about 1150 to 1370 (for Norway) and from 1150 to 1540 (for Iceland). The texts are from before the printing era and are thus preserved in manuscripts, most of which have been edited and published in scholarly editions.
In the decades following the establishment of the dictionary, the staff was mainly concerned with building the foundation for a print publication. The final result was an archive consisting of around 750.000 handwritten slips, organized under ca. 75.000 lemmas.
ONP keeps track of a variety of different information about each headword, and its quoted citations; more so than most dictionaries. The level of complexity and orthographic details is quite high.
The actual printed publication began with an index volume (ONP Registre 1989), which contained information on textual sources, editions and manuscripts of all the material ONP had collected. Three more volumes (ONP 1-3, covering the alphabet from *a* to *em*), came out between 1995 and 2004, when the print publication was put on hold and it was decided that future publication of the dictionary was to be on an electronic platform, available online. After the necessary preparation work, which including scanning the paper slips, in 2010 the first version of ONP Online was published on the web, at *onp.ku.dk*.

ONP Online brings together three basic types of dictionary entries, i.e. entries from the printed volumes, later dictionary entries edited online, and the ‘bare’ entries consisting only of a headword plus scanned citation slips. All three types of entries contain citations, which are connected through the reference sigla to a scanned page of the actual scholarly editions. The online version of the dictionary is dynamic and is continually expanding and improving, since new material and features are added as the editing work progresses.

The database and editing system

The reason why ONP has a relational database system at its core has deep historical roots. Already in the 1980s, ONP started to use its own digital word lists and data organization programs. ONP has since these early days had its own tailor-made editing software, developed over a long period by the dictionary editor and programmer Bent Christian Jacobsen (d. 2014). The software has evolved with the project and has continuously been amended and improved upon, adapting to new technological requirements and editorial needs. The result is a dictionary editing system that is optimally equipped to handle the complexity of the source material and transform the detailed information about each word into an accessible dictionary entry.

In its current configuration, the ONP system has at its core an Oracle database, which consists of a variety of tables, each relating to a specific aspect of the dictionary data. These tables are then interlinked in various ways. The programs to access the information in the database are currently written in Delphi. This reflects the historical origin of the system, going back to the 1980s Pascal programming language. This system, along with some secondary programs, provides an encompassing dictionary editing and publishing environment.

The database tables contain a multitude of fields with information related to each entry and each citation, e.g. type of word, keyword for phrases and collocations, homograph number (if needed), inflected form of the headword, headword, definition number, editorial commentary, citation, commentary, reference to cited work, edition reference, page/line reference, slip number (assigned to each slip in the archive), subentry number (when more than one entry is needed to accommodate the amount of information). Additional fields contain the information about other relevant factors, such as variant readings, foreign parallels, dating of the manuscripts, geographical provenience (Norway or Iceland) and various other kinds of information. All in all there are more than 200 fields associated with each entry.

Final words

We feel that our interests and objectives fit well with the overall goals of the Master Class. We are excited about the prospect of learning more about using xml when editing and managing lexicographic data and how to use that knowledge to improve our dictionary. We look forward to spending a professionally stimulating week in Berlin and hope to hear from you soon.

Best regards,

Ellert Thor Johannsson,
Dictionary Editor,
A Dictionary of Old Norse Prose, Copenhagen,
email: nkv950@hum.ku.dk

My name is Simonetta Battista, I'm Italian and I have been living in Denmark since 1990. I am one of four editors at the Dictionary of Old Norse Prose (ONP) in Copenhagen. My original field of interest is the hagiographic literature which was translated from Latin into Old Norse in the Middle Ages. I have a Ph.D. degree with a dissertation about "Translation or redaction in three Postola Sögur" (Sagas of the Apostles). In my research I have studied the source material of Saints' Lives to see how the bulk of Latin literature introduced with the Christianization of Scandinavia has been rendered in Old Norse.

I have been an editor at ONP since 2010, but I have been involved in the project since 1993. In my lexicographic work I have contributed to various aspects of the process of adapting the lexicographic data to the current online digital edition of the dictionary.

In recent years I have worked on presenting the ONP dictionary for the scholarly community and together with my colleague Ellert Johansson have attended several lexicographic conferences. I would like to expand my horizon by gaining new knowledge and insight into methods and techniques for the creation, management and use of digital lexical data. My colleague and I would like to attend the course together because we believe that our approaches are complementary and we are used to close collaboration in our lexicographic work. I hope therefore that our applications will be taken into consideration.

Why we want to attend this course

What we want to focus on during this Masterclass is how to implement some of the ideas we have for improving our dictionary. ONP is characterized by a large collection of citations and very detailed information about the material. Currently we work with the dictionary material in a database environment where the data are arranged in a variety of different interlinked tables and accessed through different sql-queries. Although we are used to working with our data in this way we are very much aware of how xml is being used in lexicographical work and serves as the basis of many popular editing systems. Therefore we are eager to increase our knowledge about the possibilities of its use.

Besides our general interest in learning more about the xml-based approach to dictionary structuring and editing we have two main interests. Firstly we would like to learn about ways to integrate xml-marked up data into our dictionary system. The main reason is that in recent years the amount of available electronic scholarly editions of Old Norse texts has increased tremendously and many of them are encoded in xml. Furthermore a good number of these editions include lemmatization of the text which makes them optimal to use in a dictionary context. Some texts are even freely available online, such as the texts found in the Medieval Norse Text Archive (menota.org). We would like to be able to extract such lemma information from the xml-files of the editions and arrange them in a dictionary entry structure, similar to the one we use for the ONP. We would also like to create a link between such an xml-created structure and the lemma list from our dictionary. Integrating such resources into ONP Online would provide additional citations to structured dictionary entries, as well as a wider variety of searchable examples of word use.

Secondly we would like to gain better insight into some of our dictionary data. The majority of the citations found in our database are a result of selective excerpting of scholarly text editions. We do not know how representative they are of the vocabulary of particular texts or periods, as the excerpting was only governed by the individual editors' judgment. We would like to figure out how

the vocabulary of Old Norse is represented in the ONP database and which words have been selected in the early excerption process. In order to facilitate this research we would like to analyse freely available xml-files containing lemmatized non-scholarly editions of some essential saga-texts and compare them to the corresponding citation material found in the ONP database.
The material we are bringing with us:

- Two fully lemmatized Old Norse texts in xml-form: *Strengleikar* (38453 words), *Barlaams saga* (76411 words), which are freely available in the Medieval Nordic Text Archive (menota.org).
- Complete ONP wordlist in table (and potentially xml) format
- Two saga texts, *Grettis saga* and *Fóstbræðra saga* in xml, normalized to Modern Icelandic and lemmatised using automated linguistic analysis (92.7% accuracy).
- Subset of citations from ONP in table format (and potentially xml) from those two sagas.

We, the editors of ONP, believe that by participating in this course we would gain important insight into the xml-based tools available for lexicographic work and would receive valuable training working with creating, managing and using digital lexical data. We are sure this experience would benefit the ONP dictionary in the long run and also enable closer work with the Arnamagnæan Institute's current focus on digital scholarship, as is evident by such xml-based projects as *Script and Text in Time and Space* and the *Stories for all time: The Icelandic Fornaldarsagas*, spearheaded by Prof. Matthew Driscoll.

Background of the ONP project

In what follows is a brief historical overview of the project to show where we are coming from. This overview puts the project in context and helps explain the nature of the ONP dictionary data, the development of the dictionary through the years and the current state of affairs.
The project was established in 1939 as a new, historical, scholarly dictionary which should provide an original lexical description of Old Norse with a large number of examples of word use and various kinds of supplementary secondary information.
The source material consists of Icelandic and Norwegian medieval texts, from about 1150 to 1370 (for Norway) and from 1150 to 1540 (for Iceland). The texts are from before the printing era and are thus preserved in manuscripts, most of which have been edited and published in scholarly editions.
In the decades following the establishment of the dictionary, the staff was mainly concerned with building the foundation for a print publication. The final result was an archive consisting of around 750.000 handwritten slips, organized under ca. 75.000 lemmas.
ONP keeps track of a variety of different information about each headword, and its quoted citations; more so than most dictionaries. The level of complexity and orthographic details is quite high.
The actual printed publication began with an index volume (ONP Registre 1989), which contained information on textual sources, editions and manuscripts of all the material ONP had collected. Three more volumes (ONP 1-3, covering the alphabet from *a* to *em*), came out between 1995 and 2004, when the print publication was put on hold and it was decided that future publication of the dictionary was to be on an electronic platform, available online. After the necessary preparation work, which including scanning the paper slips, in 2010 the first version of ONP Online was published on the web, at *onp.ku.dk*.

ONP Online brings together three basic types of dictionary entries, i.e. entries from the printed volumes, later dictionary entries edited online, and the ‘bare’ entries consisting only of a headword plus scanned citation slips. All three types of entries contain citations, which are connected through the reference sigla to a scanned page of the actual scholarly editions. The online version of the dictionary is dynamic and is continually expanding and improving, since new material and features are added as the editing work progresses.

The database and editing system

The reason why ONP has a relational database system at its core has deep historical roots. Already in the 1980s, ONP started to use its own digital word lists and data organization programs. ONP has since these early days had its own tailor-made editing software, developed over a long period by the dictionary editor and programmer Bent Christian Jacobsen (d. 2014). The software has evolved with the project and has continuously been amended and improved upon, adapting to new technological requirements and editorial needs. The result is a dictionary editing system that is optimally equipped to handle the complexity of the source material and transform the detailed information about each word into an accessible dictionary entry.

In its current configuration, the ONP system has at its core an Oracle database, which consists of a variety of tables, each relating to a specific aspect of the dictionary data. These tables are then interlinked in various ways. The programs to access the information in the database are currently written in Delphi. This reflects the historical origin of the system, going back to the 1980s Pascal programming language. This system, along with some secondary programs, provides an encompassing dictionary editing and publishing environment.

The database tables contain a multitude of fields with information related to each entry and each citation, e.g. type of word, keyword for phrases and collocations, homograph number (if needed), inflected form of the headword, headword, definition number, editorial commentary, citation, commentary, reference to cited work, edition reference, page/line reference, slip number (assigned to each slip in the archive), subentry number (when more than one entry is needed to accommodate the amount of information). Additional fields contain the information about other relevant factors, such as variant readings, foreign parallels, dating of the manuscripts, geographical provenience (Norway or Iceland) and various other kinds of information. All in all there are more than 200 fields associated with each entry.

Final words

We feel that our interests and objectives fit well with the overall goals of the Master Class. We are excited about the prospect of learning more about using xml when editing and managing lexicographic data and how to use that knowledge to improve our dictionary. We look forward to spending a professionally stimulating week in Berlin and hope to hear from you soon.

Best regards,

Simonetta Battista,
Dictionary Editor,
A Dictionary of Old Norse Prose, Copenhagen,
email: sb@hum.ku.dk

Proposal for Lexical Data Masterclass

My name is Michal Boleslav Měchura, I am a PhD student in computational lexicography at Masaryk University in Brno, Czech Republic. I am the author of the Lexonomy[1] dictionary writing system and a co-author of the European Dictionary Portal.[2] I have worked on many lexical data projects in Ireland including the New English–Irish Dictionary,[3] the Dictionary and Language Library[4] and the National Terminology Database for Irish.[5]

My motivation for attending the Masterclass is to understand lexical data standards better. Even though I have spent a large section of my career processing lexical data in various encoding formalisms (including XML, relational databases and informal markdown notations) I have always used home-grown schemas, in almost complete disregard for external standards. I want to fill that gap in my knowledge. In particular, to hope the Masterclass will help me become closely acquainted with the TEI guidelines for dictionaries and with how they are typically used on real-world projects.

More specifically, I propose to use the Masterclass as an opportunity to revisit three retro-digitization projects I have been involved in:

- **Foclóir Gaeilge–Béarla,**[6] an Irish-English dictionary from 1977
- **English–Irish Dictionary**[7] from 1959
- **An Foclóir Beag,**[8] a monolingual Irish dictionary from 1991

These three dictionaries, although largely out of date, are still widely respected in the Irish-language community in Ireland. They have recently been retro-digitized and encoded (under my supervision) in various home-grown formats. I have documented the process in a blog article.[9]

As an exercise for the Masterclass, I would like to investigate how I can re-encode the dictionaries in (some version of) TEI, preferably by automatic conversion from their existing formats. The dictionaries are copyrighted (for the time being, at least) but I will obtain permission to publicly share a few sample entries, if suitable.

Michal Boleslav Měchura

Natural Language Processing Centre
Faculty of Informatics
Masaryk University
Botanická 68a
602 00 Brno
Czech Republic

valselob@gmail.com

[1] http://www.lexonomy.eu/
[2] http://www.dictionaryportal.eu/
[3] http://www.focloir.ie/
[4] http://www.teanglann.ie/
[5] http://www.tearma.ie/
[6] http://www.teanglann.ie/en/fgb/
[7] http://www.teanglann.ie/en/eid/
[8] http://www.teanglann.ie/en/fb/
[9] https://multikulti.wordpress.com/2014/01/04/how-to-retro-digitize-a-dictionary/

Lionel TADJOU TADONFOUET
2 Rue Simone IFF
75012 Paris
France
☎ *+(33) 6 56 84 98 87*
✉ *lionel.tadonfouet@inria.fr*

ORGANIZING COMMITTEE of Lexical Data Master Class
Berlin Brandenburg Academy of Sciences (BBAW)

October 17, 2017

Dear Sir(s)/Madam(s),

I'm Lionel TADJOU and I hold a Master Degree in Computer science from the University of Dschang, Cameroun (2015). My research work in Master focused on the evaluation of structural preferences queries on XML documents [1], opposed to exact query on XML document which generates as answer an empty result (case of too specific queries) or too large result (case of too vague queries). I also worked as Software engineer for the pass two years in private financial companies, designing applications for electronic payment and banking application. Currently Research engineer at the French National Institute for Computer Science and Applied Mathematic (Inria) in Paris, and member of the Almanach team (INRIA – EPHE), I work as a Web developer for the PARTHENOS project. I'm designing the Standardization Survival Kit (SSK) website which is a platform dedicated to support the modeling and management of research data for Arts and Humanities researchers. I'm efficient on Java (Spring) programming language, JavaScript(Jquery), AngularJs & Angular 2, Bootstrap, REST API and XML manipulation. I'm also Beginner in Python programming and machine learning. I would like to participate to the Masterclass (Lexical Data Master Class), in order to gain more knowledge in computational linguistic on under resourced languages.

Most of local African languages are underresourced and may disappear as shown on UNESCO Atlas of the World's Languages in Danger [1] if nothing is done to keep them alive. Underresourced languages could be defined as languages spoken by a few number of people, with few available corpora.To avoid that underressourced languages disappear, or at least to reduce the number of them dying, I propose to build a platform dedicated to help people learn these languages. This platform could be set up easily, if only those languages had many corpus written (text) or spoken (audio, video).

Therefore the first step of this project consist in focusing on a way to generate corpus, mainly by getting data from the few people speaking them by customizing or building a tool like LIG-AIKUMA [2], but also by ingesting data from lingustic archives like dictionnaries [2] or other resources [3] for each languge such as the Yemba [4] language for instance. A survey [3] also presents the recent contributions made for under-resourced languages. The idea to build an engaged web application where users would upload contents (text, expressions, images, audio) for differents languages, with some users identified as language specialists who could evaluate those contents, should also be taken into account. In this way users will be classified or grouped per valide contents uploaded and

[1] UNESCO Atlashttp://www.unesco.org/languages-atlas

[2] Yemba dictionnaryftp://ftp.cis.upenn.edu/pub/sb/papers/dictionary/dictionary.pdf

[3] Yemba resources http://eveilyemba.org/

[4] Yemba language https://en.wikipedia.org/wiki/Yemba_language

behind we will have many structured resources.

The second step will focused on building a learning web/mobile application like Duolingo [5] to learn those underresourced languages. This learn app will uses concepts as Named-entity, TreeTagger [6] tool and other Natural Languages Processing technologies. Based on languages commonalities or belonging of same sub-group of language and learner's abilities,the app will proposed content of other languages that the learner could easily take hands on.

I would like start this project by building a Lexical content for Yemba which is an underresourced language of Eastern Grassfields languages (sub-group of bantou languages). This Masterclass would be a great opportunity to launch this project, by helping me build a lexical model for languages of same sub-group or with some commonalities. Within this Masterclass I would learn about lexicon model, principals concepts of lexicon, representation of lexical content and other related resources. Because my goal is to build this platform for many languages I hope I will also learn during this Masterclass how to manage a community and digital resources linked to a project.

References

[1] Lionel Tadonfouet Maurice Tchoupe Tchendji. “Parallel Speech Collection for Under-resourced Language Studies Using the Lig-Aikuma Mobile Device App”. In: *Conference: 12th Africain Conference on Research in Computer Science*. Saint Louis, Sénégal, 2014. URL: http://www.cari-info.org/cari_2014/p35.pdf.

[2] David Blachon u. a. “Parallel Speech Collection for Under-resourced Language Studies Using the Lig-Aikuma Mobile Device App”. In: *Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*. Yogyakarta, Indonesia, Mai 2016. DOI: 10.1016/j.procs.2016.04.030. URL: https://hal.archives-ouvertes.fr/hal-01350065.

[3] Laurent Besacier u. a. “Automatic speech recognition for under-resourced languages: A survey”. In: *Speech Communication* 56.Supplement C (2014), S. 85–100. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2013.07.008. URL: http://www.sciencedirect.com/science/article/pii/S0167639313000988.

[5] Duolingo https://fr.duolingo.com

[6] TreeTagger http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Scholarly Background

Sabine Wahl MSt, M.A., is a lexicographer for the Wörterbuch der bairischen Mundarten in Österreich ('Dictionary of the Bavarian Dialects in Austria', WBÖ) at the ACDH research department 'Variation und Wandel des Deutschen in Österreich' ('Variation and Change of German in Austria'). In addition, she works as a lecturer at the Department of German Philology at the University of Vienna.

Before joining the Austrian Academy of Sciences in April 2017, she was a lecturer at the University of Bremen (German linguistics) and an academic assistant at the Chair for German Linguistics at the University of Eichstätt-Ingolstadt. In Eichstätt, she also worked as a lecturer in the Master's degree course 'InterculturAd – Werbung interkulturell' (and as a guest at the Åbo Akademi University in Turku), was involved in a DFG project on word-creation and was a researcher in a project on German and Italian brand names, carried out in cooperation with the Università degli Studi di Verona.

She studied at the Universities of Eichstätt-Ingolstadt (German, English, Spanish and Educational Studies: State exam for teachers in German, English/TESOL; M.A.: German Linguistics, Medieval German Studies and English Linguistics) and Oxford (Master of Studies in European Literature – German).

Her research interests include the phonology and morphology of words and names (esp. brand names), language change, dialectology, language and music as well as the linguistic and multimodal design of advertisements and commercials.

Specific Interest in the Field / Topics

As I am part of a new team of lexicographers for the *Dictionary of Bavarian Dialects in Austria* (see below), I am generally interested in learning more about the creation, management and publication of digital lexical data (online and print articles; online research platform). With regard to the teaching topics offered, the following aspects of e-lexicography would be highly relevant to my work:

- TEI based representations for lexical content
- Working efficiently with an XML editor
- Dealing with morphology in a digital lexicon
- Managing examples in dictionaries
- XPath for Lexicographers
- Issues related to data management (FAIR principles, DMP, sustainability)
- Managing digital lexica as online resources

Presentation of the Dataset

The *Dictionary of Bavarian Dialects in Austria* ('Wörterbuch der bairischen Mundarten in Österreich', WBÖ) is a long-term project that aims at the comprehensive lexicographic documentation of the Bavarian dialects in Austria and South Tyrol. The language data was compiled during the 1st half of the 20th century either indirectly with the help of so-called collectors ("Sammler") on the basis of questionnaires ("Fragebücher"), or directly during field explorations, and was further complemented with excerpts from specialised literature. All data was written down on paper slips, which represent the so-called main catalog ("Hauptkatalog") with approximately 3.6 million entries. To date, five volumes of the WBÖ have been published, covering the entries *A-Ezzes.*

To facilitate and accelerate the process of writing dictionary articles, the information on the paper slips was entered manually into a TUSTEP system in the 1990's and, subsequently, converted into TEI-XML.

The TEI-XML data is the basis for our work on a revised and modernised conception of the dictionary. In addition to the publication of "classical" dictionary articles (including the etymology, the phonetic variants, and the geographic distribution of each headword), we are planning to provide a web-based research and information platform on the lexis of German in Austria. For this purpose, the digitized data of the main catalog as well as other materials (e.g., scans of the paper slips and questionnaires) are going to be integrated into the online platform. This will enable users to conduct extensive research on the material (e.g., geographic distribution words visualised on maps; onomasiological queries, i.e. different linguistic forms connected with the same semantic concept).

The data structure that will be used for the dictionary and the research platform will also be TEI. This will involve the creation of a complex set of entry templates in order to accommodate the various different data fields common in dictionary articles, especially in dialect dictionaries. This usage will be rather novel as it is less common for a print dictionary to be compiled, generated and edited first in TEI. By developing this set of templates, more general issues of (e-)lexicography are addressed: the structural and content demands of print dictionaries versus those of digital dictionaries (plus a web-based research platform) and the need for a user-friendly software editing tool for non-experts in order to edit and create articles in TEI directly.

Bowers, J., & Stöckle, P. (2017). “TEI and Bavarian dialect resources in Austria: updates from the DBÖ and the WBÖ”. Submitted for publication.

Application to LexMC

Name: Maria Koutsombogera

Affiliation: Trinity College Dublin, School of Computer Science and Statistics

Research Group: Computational Linguistics

E-mail address: koutsomm@scss.tcd.ie

Telephone: +353838836149

Joint project with Anna Vakalopoulou, "English-Greek Dictionary"

Background

I am currently a Marie Curie research fellow at Trinity College Dublin. From 2002 to 2015 I was affiliated with the Institute of Language and Speech Processing in Athens, Greece. I hold a PhD in Linguistics (University of Athens, 2012), a MSc in Language Technology (School of Electrical and Computer Engineering, NTUA & Department of Linguistics, UoA, 2004) and a bachelor's degree in Linguistics (University of Athens, 2000).

My research interests focus on the development and processing of textual and multimedia corpora, multimodal conversation analysis, and computational morphology, syntax and semantics.

Throughout my research experience I have been involved in several lexicographic and corpus development tasks, including the following:

- Development of computational morphological lexica
- Development of resources for dependency parsing
- Development of resources for fact extraction from texts
- Digital dictionaries editing
- Thesauri enrichment
- Discourse annotation
- Dialogue act annotation
- Development of resources for modeling sensorimotor experiences
- Development of resources for multimodal communication processing
- Use of several annotation editors employing xml schemata

Interest for the field and topics to explore

I believe that I would benefit from my attendance to the masterclass in terms of improving the understanding about creating, managing and using digital lexical data. Specifically, I am interested in the identification and use of XML editors and getting involved in the hands-on lab sessions. Similarly to this, I am also interested in interoperability issues and mappings between different approaches. Moreover, I would like to learn more about managing digital lexical as online resources. I also think that it's a unique opportunity to meet and interact with experts in the field.

Moreover, I would like to look into the applicability of tools and methods to the project we are implementing. Particularly to this dictionary project that we're working on, I would like to identify the appropriate solution for using a standards-compliant tool, to evaluate the methodology we use, and to consult with the experts about challenging issues that we're facing and to discover solutions about modeling and enriching our resource.

Presentation of the dataset to be created/ encoded/ enriched

This is a joint project with applicant Anna Vakalopoulou.

The dataset we plan to work with is an English-Greek Dictionary. It currently consists of 17,000 lemmas and is in text editor format.

The dictionary comprises both single- and multi-word entries. Lemma differentiation is based principally on morphology as usual in dictionary making practice, followed by part of speech. Sub-lemmas include derivatives with no special lexicographic interest, i.e. the meaning and/or usage of which is straightforward and needs no further explanation.

Information contained for each lemma includes the following:

- phonetic transcription,
- alternative lemmas,
- senses, and
- sub-lemma(s).

Information contained for each sense includes the following:

- syntactic collocates
- part of speech
- irregular types
- language variety
- register
- topic
- translation into Greek
- example(s) & translation into Greek
- expressions/idioms & translation into Greek
- phrasal verbs

Below is a sample entry of the verb ***joke***.

joke /δΖ↔Υκ/ [**about**, **with**] *v i*, αστειεύομαι, κάνω πλάκα ● *Don't joke with him on religious matters*, μη του κάνεις πλάκα για θρησκευτικά θέματα ● *It's not wise to joke about such a serious matter*, δεν είναι φρόνιμο να αστειεύεσαι με τόσο σοβαρά θέματα. **IDM & EXP joking apart / aside** [*BrE*] πέρα από την πλάκα, χωρίς πλάκα ● *Joking apart, he looked great yesterday!* Χωρίς πλάκα, ήταν πολύ ωραίος χθες! **"Only joking!"** 'Απλώς αστειεύομαι!'. **"You must be joking! / You're joking! / You've got to be joking!"**, 'Ασφαλώς θα αστειεύεσαι!' ● *You'll do that? You must be joking!* Θα κάνεις τέτοιο πράγμα; Ασφαλώς θα αστειεύεσαι!

We would like to structure the dictionary in a way that enables better display and storage and to identify ways to better represent the content, specifically:

(a) Convert the source files (text editor content) to xml.
(b) Select appropriate xml editors.
(c) Create syntactically correct xml files and manage format challenges, such as character encoding.
(d) Handle links to other resources or complementary files.
(e) Specify the morphology representation.
(f) Ensure correct format and consistency.
(g) Enrich current content with more lemmas and updated examples.
(h) Create a digital content representation with the appropriate structure resulting in a user-friendly environment.

The dataset I will bring is that from my language documentation work which is part of my PhD project from the Mixtepec-Mixtec language (MIX).

The core data set includes the following:

- a collection of time aligned TEI utterance files (originally transcribed in Praat and converted via XSLT) from recordings of consultation sessions with native speakers

- content collected from personal communications with speakers and another researcher

- a collection of TEI versions of children's texts in the language's working orthography published by the Summer Institute of Linguistics (SIL) Mexico

- TEI feature structures for phonology as well as morphosyntactic and semantic features of the language

- TEI encoded (mostly only partially at present) of a small number of published literature on MIX covering description of the morpho-phonology of the language

- TEI encoded personal journal written over the course of ten days by a speaker consultant

- I currently have a manually compiled dictionary, but a goal is to generate a new TEI dictionary from the corpus contents using XSLT

Additionally, there are a great deal of metadata contained in this dataset that have yet to be entirely dealt with. With the exception of the TEI dictionary, the rest of this data is being annotated using standoff annotation.

The MIX content comes with a number of complications due to several different factors, one of which is the fact that the standard orthography is still under development (being developed by SIL Mexico in collaboration with the Academy for the Mixtec Language). Thus some of the older published SIL documents do not match the newer ones, and even the newer publications do not match the most current writing system used by the SIL researchers. Also, speakers do not use this system when they produce written material and thus in addition to the variation we get from the published sources, which is at least somewhat consistent in the temporal timespan in which certain spelling conventions were used, this data set also contains written materials in MIX from speakers who often use their own spelling practices which are so different, it is necessary to create a entirely separate normalized version of the texts then which must be linked to the source using pointers.

At present I have for the most part found solutions to each of these issues within the TEI but I would be interested in seeing if there are any ways in which I may improve upon my adopted approach either in the means of encoding. There also remain a large number of small things that I would like to discuss with other experts.

As for the sessions, I particularly would like to attend sessions concerning: Managing digital lexicons as online resources; Presentation of dictionary content using XSLT; Dealing with morphology in a digital lexicon. I have significant interest as well experience in every other subject on the topic list, I would be glad to contribute what I can to help others in the sessions and I am always looking to improve my knowledge of all aspects of my field.

What I would like to gain from participation is to possibly find some solutions or at very least some recommendations for improvements to certain aspects of my encoding methodology for the issues mentioned above, the workflow and ways in which I may lighten the manual workload through usage of XSLT or other scripting processes to annotate and/or re-format aspects of the data.

Creation a data model for encoding the *Old Slavonic – Greek / Latin Dictionary* (*Lexicon Palaeoslovenico–Graeco-Latinum*) published in 1865 by F. Miklošič in TEI

dr. sc. Martina Kramarić
Research assistant
Department of the History of Croatian Language and Historical Lexicography
Institute of Croatian Language and Linguistics
Republike Austrije 16, HR − 10000 Zagreb, Croatia
t: +385 1 3783 811
f: +385 1 3783 803
cell. +385 95 907 79 95
mkramar@ihjj.hr / www.ihjj.hr

Scholarly background

I am a research assistant in the Department of the History of Croatian Language and Historical Lexicography in the Institute of Croatian Language and Linguistics. My study (finished Postgraduate Study of Croatian language history and dialectology, thesis: Czech loanwords in Zrcalo člověčaskogo spasenja (1445) in the context of Old Croatian translations from the Old Czech language (February 18th, 2015)) and professional experience in the field of the history of the Croatian language gave a strong theoretical background for understanding this complex old dictionary. I am working with old manuscripts and documents (Latin, Cyrilic and Glagolitic), some of them written in the Old Church Slavonic language, and I am familiar with the old scripts and sources used to create dictionary. The experience in the field of retro-digitization and preparation of digital editions of the old dictionaries I have gained on the various workshops and trainings:

- ENeL TRAINING SCHOOL 2015: Standard tools and methods for retro-digitising dictionaries, 6 – 10 July 2015, Lisbon, Portugal

- WG2 Meeting / Workshop, COST ENeL WG2 meeting, Barcelona, Spain, 31 March – 1 April 2016

- Short term scientific mission at the Belgrade Center for Digital Humanities, working with Toma Tasovac, Belgrade, Serbia, 25 – 29 September 2017

I have participated in the Cost action, ENel, WG 2 Retro-digitized dictionaries. During my stay at National Library of the Czech Republic, Manuscriptorium department in the frame of the Cendari Visiting Research Fellowships Programme (from August 10th till October 31th, 2014) I worked on project *Creating parallel mediaeval corpora: a database of a TEI-encoded version of the Old Croatian text Zrcalo and its Old Czech templates*, where I was introduced in the usage of TEI in Digital Humanities.

About project

In my project I would like to present the guidelines for creation a data model for encoding the Miklošič dictionary *Old Slavonic – Greek / Latin Dictionary* (*Lexicon Palaeoslovenico–Graeco-Latinum*) in TEI. The goal is to create a digital edition of the Miklošič dictionary. The project of digitization of Miklošič dictionary is developed in cooperation of the Belgrade Center for Digital Humanities and Institute of Serbian Language of the Serbian Arts and Sciences Academy for the dictionary platform Raskovnik.org. (http://raskovnik.org).

About dictionary

Miklošič's dictionary *Old Slavonic – Greek / Latin Dictionary* (*Lexicon Palaeoslovenico–Graeco-Latinum*) (1862–1865) although written a century and a half ago, is still the most relevant dictionary in diachronic Slavic studies. The dictionary consists of 1,172 pages with approximately 42,000 entries. In this dictionary, Old Slavonic word meanings are provided in Greek and Latin, and are compared with words from all European languages, taking into account all Indo-European languages relevant in comparative grammar. It descriptions take sources from all Slavic regions into account. Etymologies of the words are also provided in some entries. This is the only dictionary that embraces both words from the canonical Old Slavonic language and words from all Old Church Slavonic redactions. The sources (consisting of a total of 280 published and unpublished written sources in all Church Slavonic redactions and the canonical Old Church Slavonic language) used date from between the 11th and 17th centuries.

The headword (lemma) and the examples from the sources are written in the canonical Old Slavonic form and in Cyrillic script. Grammatical descriptions are provided in Latin script, as are the Latin meanings. Greek meanings are written in the Greek alphabet.

Thanks to its wide scope, its rich base of sources and etymological data, and its large number of entries, this dictionary is an extremely valuable resource of lexical data for all scholars interested in Slavic comparative studies, etymology, and Slavic philology in general. Therefore, creation of its digital edition is much needed and very important.

The first step in the process of the creation of the digital edition of the printed dictionary is the semantic segmentation of each dictionary entry. In the case of Miklošič's dictionary each dictionary entry consists of a headword, a simple grammatical description (parts of speech, but for verbs for example without information on verb aspect). Greek and Latin meanings are provided: Greek ones are provided for words excerpted from sources translated from Greek, and Latin ones are provided for all words. Definitions are provided through Latin and/or Greek meanings. Sometimes the note "*sensus nobis ignotus*" or "*vox obscura*" is written in place of equivalents. In a number of entries, usage examples of the headword are provided from the textual sources, and if the text has been published, the page of the given example is quoted. The end of the entry provides word equivalents in other Slavic or other Indo-European languages. Hungarian and Romanian Slavic loanwords are also mentioned in some dictionary entries. Usage examples are provided by quoting collocations, phrases, or sentences in which the headword (lemma) is used. Some entries contain a variety of examples, phrases, etc. Complex dictionaries like Miklošič's *Old Slavonic – Greek / Latin Dictionary*, which involve multiple languages without providing a real definition of the lemma (meanings are provided through Latin and Greek translations) and whose dictionary entries are structured differently, involve a great deal of preparation work in segmentation and structural definition. The proper structure of a digital dictionary must be based on the precise semantic annotation of every single piece of linguistic data within the dictionary entries. So, before creation of guidelines the work on structuring and pre-defining of dictionary entries will be presented presenting infrastructural challenges in that process. In this process the best practices in the retro-digitization and preparation of digital editions for the dictionary platform raskovnik.org were used as a role model.

Miklošič's dictionary is very complex one and and its lack of a unique and standardized lexicographic form for every dictionary entry makes it even more difficult and demanding to encoding. Despite that, the final result would be of great value to all scholars, and would serve as an excellent general example for the future digitization of any kind of dictionary.

Application to LexMC

Name: Anna Vakalopoulou

Affiliation: Institute for Language and Speech Processing / Athena RC (ILSP)

Research Group: Computational Lexicography

E-mail address: avacalop@ilsp.gr

Telephone: +302106875447

Joint project with Maria Koutsombogera, "Engish-Greek dictionary"

Background

I have been a practicing lexicographer for 20 years and I currently work at ILSP as a scientific associate. I hold an MA in Lexicography (University of Exeter, UK, 1997).

My research interests focus on the development and processing of monolingual and multilingual dictionaries and other reference works for human and machine use.

I was editor in chief in the proposed work.

I have been involved in several lexicographic tasks as an editor, senior editor including:

PROJECT	ROLE	YEAR(S)
Collins COBUILD Student's Dictionary	Internship	1996
A Thematic Dictionary of Greek	Editor	1997-1998
My First Dictionary for Schoolchildren	Sole editor	1999-2000
NOEMA: Greek Sign Language dictionary	Editor	2000
Xenion Lexicon for Tourists (EL-EN)	Editor	2000-2001
EL-TR dictionary for children in *A Knight in the* Castle *of Letters*, a method for teaching Greek	Editor of the Greek part	2000-2001
Greek-Turkish Dictionary for Children	Sole editor of the Greek part	2002-2004

Trilingual Dictionary of Terminology (EL, EN, TR)	Associate scientific coordinator	2003-2004
Tomi Gold 2004, electronic multimedia encyclopedia	Editor of Language & Linguistics entries	2003-2004
Diolkos, GSL-EL-EN dictionary of computer terms	Editor	2004
Lexipaideia, EL high school dictionary	Editor	2005-2010
The New Dimitrakos Dictionary, EL general language	Editor	2010-2015
	Editor in chief	2016-2017
e-MiLang, a set of 9 bilingual dictionaries for immigrants	Editor in chief	2011-2013
English-Greek Dictionary (described below)	Editor in chief	2003-2013
EPAL, EL vocational high school dictionary	Editor in chief	2013
Polytropon, SGL-EL high school dictionary	Editor in chief	2013-2014
P61, SGL-EL high school dictionary	Editor in chief	2014-2016

Presentation of the dataset to be created/ encoded/ enriched

The dataset my partner and I plan on working with is an English-Greek Dictionary, currently consisting of 17,000 lemmas in text editor format.

The dictionary comprises both single- and multi-word entries. Lemma differentiation is based principally on morphology as usual in dictionary making practice, followed by part of speech. Sub-lemmas include derivatives with no special lexicographic interest, i.e. the meaning and/or usage of which is straightforward and needs no further explanation.

Information contained for each lemma includes the following:

- phonetic transcription,
- alternative lemmas,
- senses, and
- sub-lemma(s).

Information contained for each sense includes the following:

- syntactic collocates
- part of speech
- irregular types

- language variety
- register
- topic
- translation into Greek
- example(s) & translation into Greek
- expressions/idioms & translation into Greek
- phrasal verbs

Sample pages of this work can be found in the section Supplementary Data.

Our particular interests in connection to LexMC are to find ways that will help us structure the dictionary in a way that enables better processing, storage, and display of the content, specifically:

(a) Convert the source files (text editor content) to xml
(b) Select appropriate xml editors
(c) Create syntactically correct xml files and manage format challenges, such as character encoding.
(d) Handle links to other resources or complementary files
(e) Specify the morphology representation
(f) Ensure correct format and consistency
(g) Enrich current content with more lemmas and updated examples
(h) Create a digital content representation with the appropriate structure resulting in a user-friendly environment

Application to LexMC

Name: Anna Vakalopoulou

Affiliation: Institute for Language and Speech Processing / Athena RC (ILSP)

Research Group: Computational Lexicography

E-mail address: avacalop@ilsp.gr

Telephone: +302106875447

Joint project with Maria Koutsombogera, "Engish-Greek dictionary"

Background

I have been a practicing lexicographer for 20 years and I currently work at ILSP as a scientific associate. I hold an MA in Lexicography (University of Exeter, UK, 1997).

My research interests focus on the development and processing of monolingual and multilingual dictionaries and other reference works for human and machine use.

I was editor in chief in the proposed work.

I have been involved in several lexicographic tasks as an editor, senior editor including:

PROJECT	ROLE	YEAR(S)
Collins COBUILD Student's Dictionary	Internship	1996
A Thematic Dictionary of Greek	Editor	1997-1998
My First Dictionary for Schoolchildren	Sole editor	1999-2000
NOEMA: Greek Sign Language dictionary	Editor	2000
Xenion Lexicon for Tourists (EL-EN)	Editor	2000-2001
EL-TR dictionary for children in *A Knight in the* Castle *of Letters*, a method for teaching Greek	Editor of the Greek part	2000-2001
Greek-Turkish Dictionary for Children	Sole editor of the Greek part	2002-2004

Trilingual Dictionary of Terminology (EL, EN, TR)	Associate scientific coordinator	2003-2004
Tomi Gold 2004, electronic multimedia encyclopedia	Editor of Language & Linguistics entries	2003-2004
Diolkos, GSL-EL-EN dictionary of computer terms	Editor	2004
Lexipaideia, EL high school dictionary	Editor	2005-2010
The New Dimitrakos Dictionary, EL general language	Editor	2010-2015
	Editor in chief	2016-2017
e-MiLang, a set of 9 bilingual dictionaries for immigrants	Editor in chief	2011-2013
English-Greek Dictionary (described below)	Editor in chief	2003-2013
EPAL, EL vocational high school dictionary	Editor in chief	2013
Polytropon, SGL-EL high school dictionary	Editor in chief	2013-2014
P61, SGL-EL high school dictionary	Editor in chief	2014-2016

Presentation of the dataset to be created/ encoded/ enriched

The dataset my partner and I plan on working with is an English-Greek Dictionary, currently consisting of 17,000 lemmas in text editor format.

The dictionary comprises both single- and multi-word entries. Lemma differentiation is based principally on morphology as usual in dictionary making practice, followed by part of speech. Sub-lemmas include derivatives with no special lexicographic interest, i.e. the meaning and/or usage of which is straightforward and needs no further explanation.

Information contained for each lemma includes the following:

- phonetic transcription,
- alternative lemmas,
- senses, and
- sub-lemma(s).

Information contained for each sense includes the following:

- syntactic collocates
- part of speech
- irregular types

- language variety
- register
- topic
- translation into Greek
- example(s) & translation into Greek
- expressions/idioms & translation into Greek
- phrasal verbs

Sample pages of this work can be found in the section Supplementary Data.

Our particular interests in connection to LexMC are to find ways that will help us structure the dictionary in a way that enables better processing, storage, and display of the content, specifically:

(a) Convert the source files (text editor content) to xml
(b) Select appropriate xml editors
(c) Create syntactically correct xml files and manage format challenges, such as character encoding.
(d) Handle links to other resources or complementary files
(e) Specify the morphology representation
(f) Ensure correct format and consistency
(g) Enrich current content with more lemmas and updated examples
(h) Create a digital content representation with the appropriate structure resulting in a user-friendly environment

Standardising variation: A <concept> for the *Atlas Linguistique de la France ?*

Susanne Alt
Chargée de Recherche CNRS actuellement en détachemant
susanne-alt@t-online.de

The project I would like to submit to the Lexical Data Master Class relates to a standardized encoding proposal for French dialectological data. I am familiar with the maps of the Atlas Linguistique de France since my undergraduate studies in Roman and German linguistics at the Universities of Berlin and Grenoble, including dialectology, phonetics and historical linguistics. After a PhD thesis in Computational Linguistics, I started to work as a tenured CNRS researcher on converting and encoding traditional lexical data in order to make them available for natural language processing purposes. My focus was to test ongoing standardization proposals, such as LMF and TEI, against the richness of traditional lexicographical data. Since 2007, I've been detached from CNRS to an international organisation, where I had the occasion to get acquainted with use of lexical data in real-life NLP applications through the examination of patent applications from major industrial players in the field. This experience shows that whilst access to open and standardized lexical resources remains crucial to the development of intelligent speaking agents, any sustainable encoding standard should rely on a careful and theoretically sound analysis of the underlying data. In line with my work on encoding etymological data (Salmon-Alt, 2006[1]), I would therefore like to take up the occasion of the Master Class to work on dialectological data with the aim of coming up with a proposal which may benefit a recently started project on digitizing the Atlas Linguistique de France (Gally et al., 2013[2]).

1 What is the *Atlas Linguistique de la France* ?

The *Atlas Linguistique de la France* has been edited by Jules Gilliéron and Edmond Edmont, who collected, between 1888 and 1892, phonetic, lexical and syntactic variants of dialectal forms in the gallo-roman linguistic area. Following the tradition of the first German language atlases compiled a few years ago, it is based on interviews using a questionnaire as stimulus, and materializes as a collection of 1920 geographical maps of the gallo-roman linguistic area, each of the maps identifying, for 639 selected survey points, the local counter-part of a stimulus presented to the speaker as a French isolated token, a phrase or a whole sentence. Figure 1 shows a snippet of the map n°13, wherein the French word AIGLE was realized, for example at survey point 267 as « ȅgl ». The phonetic transcription follows the phonetic alphabet created in 1887 by the French phonetician Rousselot and Gilliéron himself.

A first ALF digitization project made publicly available image scans of the maps[3]. In addition, the survey points have been geocoded and exploratory work has been initiated based on fifteen simple word maps towards a data model in view of integration into a geographical information system (Gally et al. 2013, Davoine et al, 2015[4]). The outcome is a relational data model which however appears to suffer from the following drawbacks: (i) it is a proprietary format database overlooking any of well-established standards for encoding linguistic data; (ii) it mixes up diplomatic and interpretative features without proper attribution of authorship, and (iii) it misses genericity, *inter alia* by reducing potentially complex features to simple attributes or hard-coded cardinalities. The scope of the current proposal is to overcome these drawbacks by the attempt to map a

[1] Alt, Susanne (2006). Data structures for etymology: towards an etymological lexical network. BULAG, Presses Universitaires de Franche-Comté, vol. 31, p. 1-12.

[2] Gally S., Chauvin C., Davoine P.-A., Demolin D., Contini M. (2013), "GéoDialect : Exploration des outils géomatiques pour le traitement et l'analyse des données géolinguistiques" in Géolinguistique 14, Grenoble : ELLUG, 186-208.

[3] http://cartodialect.imag.fr/cartoDialect/accueil.

[4] Davoine P.-A., Gally S., Garat P., Chauvin C., Čopi O., Cavalière C. (2015). New approach to explore and to study cartographical heritage in dialectology: application to the Linguistic Altas of France, ICC Conference.

linguistically sound analysis of the underlying data to existing standards for encoding lexical ressources.

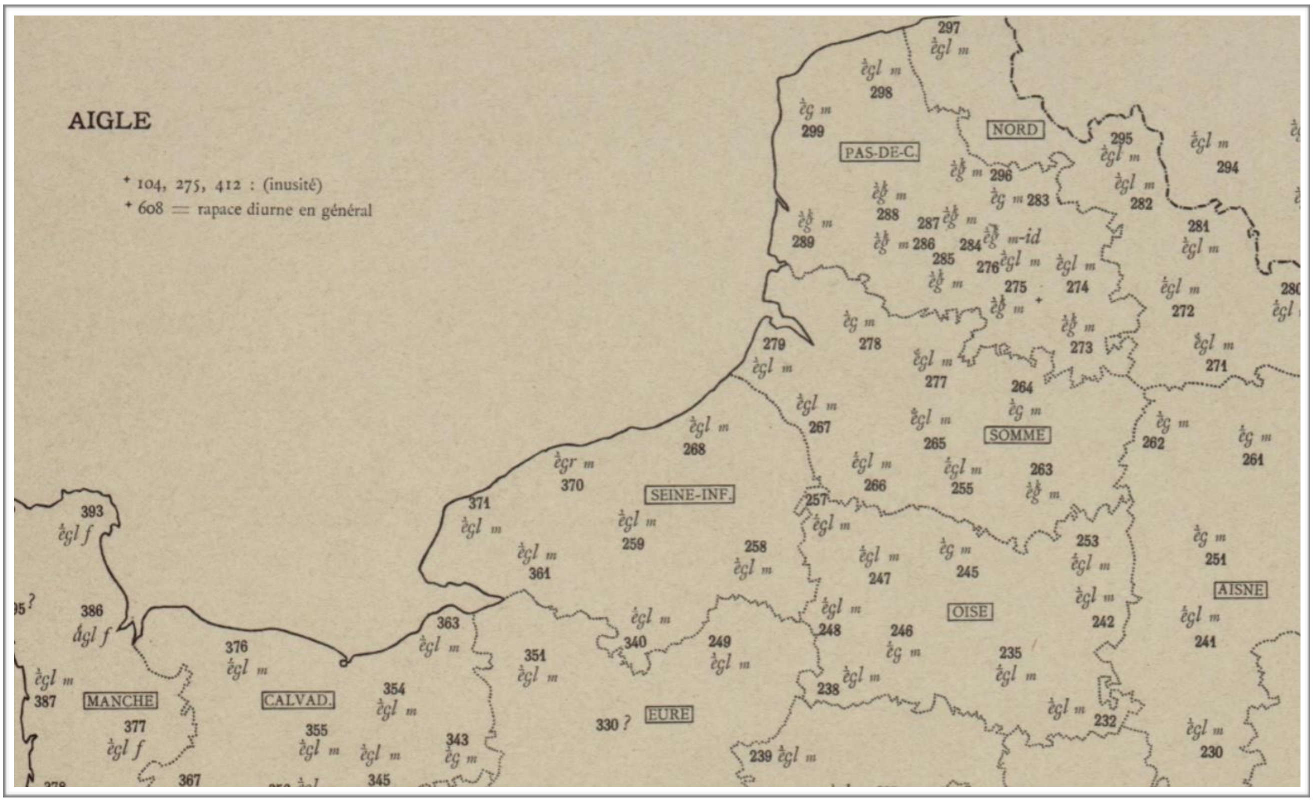


Fig. 1: ALF - snippet of the map « AIGLE »

2 Is there a need for a genuine onomasiological data model ?

Geo-linguistic data collections, together with terminological databases, are most often qualified as prototypical examples of an onomasiological lexicographical approach. The principle underlying the ALF was indeed to collect, for a given meaning, the corresponding names, cf. Greek ὀνομάζω « to name ». As opposed to a semasiological approach, an onomasiological entry does therefore not – at least in theory – associate one name with multiple meanings, but one meaning with multiple names. An abstract view of the map AIGLE would therefore be an entry associating the meaning of French AIGLE with 639 names.

For a number of extra-linguistic reasons summarized in Romary (2014)[5], encoding standards for onomasiological and semasiological lexical ressources have taken divergent ways. To a lesser extent they still continue to do so, as shows a recent proposal to enrich the TEI with a chapter genuinely dedicated to onomasiological ressources (Romary 2014, Bowers et al 2017[6], TBX in TEI[7]). I advocate here that this dichotomy – whilst surely appropriate to qualify the underlying scientific methodology – does not necessarily justify fundamentally different encoding standards and elements for the resulting lexical ressources. Based on the classical onomasiological data from the ALF, my project for the Master Class is to compare the upcoming onomasiological proposal (cf. references above) with the semasiological TEI Dictionary encoding guidelines[8]. In particular, I wish to examine to what extent the TEI Dictionary specification already provides the necessary basic blocks to encode geo-linguistic variation, how far tags would need to be

[5] Romary, Laurent (2014). TBX goes TEI – Implementing a TBX basic extension for the Text Encoding Initiative guidelines. Terminology and Knowledge Engineering 2014, Berlin.

[6] Bowers Jack, Pernes Stefan, Romary Laurent (2017). ConceptEntry: A TBX-based expansion of the TEI for the encoding of onomasiological and comparative lexical data. TEI 2017 Victoria, British Columbia, Canada, November 11-15.

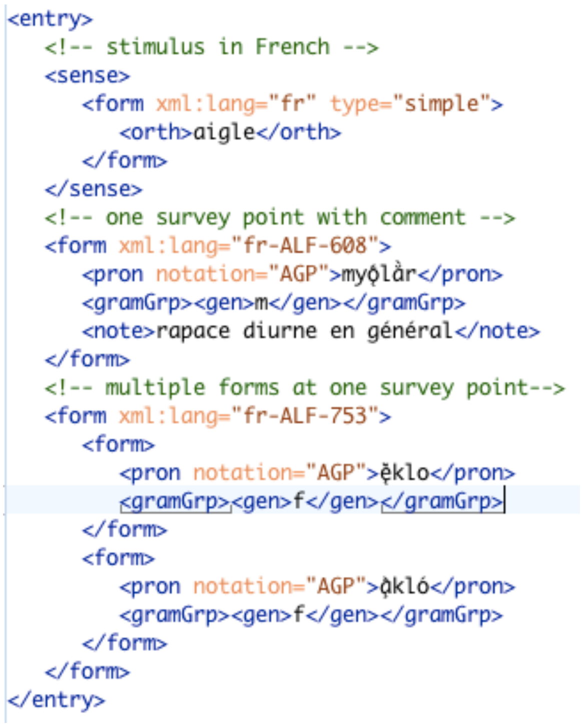
[7] https://github.com/ParthenosWP4/standardsLibrary/blob/master/terminology/specification/TBXinTEI.html

[8] http://www.tei-c.org/release/doc/tei-p5-doc/de/html/DI.html

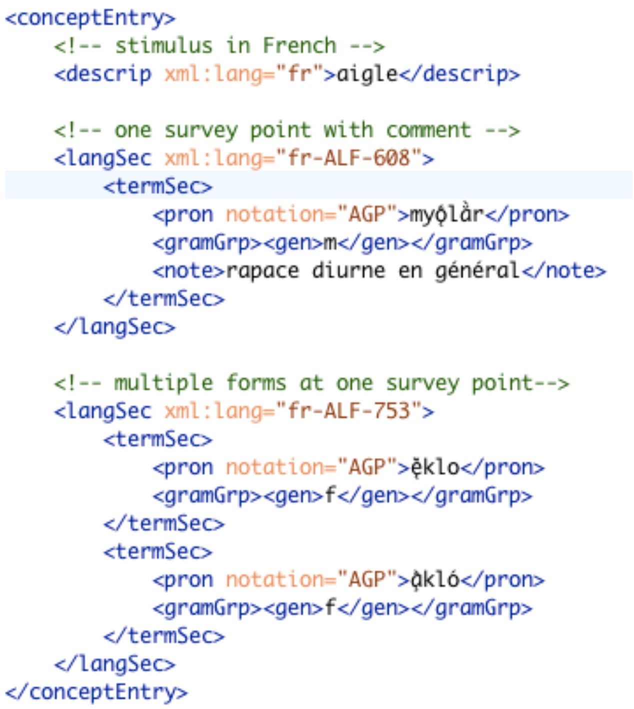
« abused », and which new elements of the onomasiological proposal would turn up to fill remaining gaps, if any.

3 Outline of onomasiological and semasiological TEI data models

Fig. 2 and 3 provide a gist of possible macro-structurations of dialectological entries. Whereas the TEI dictionary encoding (ab)-uses the traditional form-sense structure by simply inverting cardinalities, the TBX to TEI proposal uses an additional language section borrowed from TMF ISO 16624[9] in association with a « still-to-be-further-defined » concept section. During the Master Class, I would like to explore both proposal further, by taking into account the complexity of the underlying ALF data. As opposed to the simple concept AIGLE, the French stimulus may for example represent a set of morphological derivates from the same word stem, a set of semantically related concepts or a set of syntactic constructions, all of them embedded in contexts. Survey points often record several form variants for one or more input stimuli, thereby present ambiguities in respect of which stimulus they relate to and/or may overwrite the semantics of the input stimulus. In addition, a typology of notes at different levels of encoding needs to be taken into account in order to properly encode additional observations and annotations. The data therefore clearly suggest that both the concept and the form parts more often then not deviate from the simple example structures below.



```
<entry>
   <!-- stimulus in French -->
   <sense>
      <form xml:lang="fr" type="simple">
         <orth>aigle</orth>
      </form>
   </sense>
   <!-- one survey point with comment -->
   <form xml:lang="fr-ALF-608">
      <pron notation="AGP">myǫ̇lå̀r</pron>
      <gramGrp><gen>m</gen></gramGrp>
      <note>rapace diurne en général</note>
   </form>
   <!-- multiple forms at one survey point-->
   <form xml:lang="fr-ALF-753">
      <form>
         <pron notation="AGP">ẹ̆klo</pron>
         <gramGrp><gen>f</gen></gramGrp>
      </form>
      <form>
         <pron notation="AGP">ọ̀kló</pron>
         <gramGrp><gen>f</gen></gramGrp>
      </form>
   </form>
</entry>
```



```
<conceptEntry>
    <!-- stimulus in French -->
    <descrip xml:lang="fr">aigle</descrip>

    <!-- one survey point with comment -->
    <langSec xml:lang="fr-ALF-608">
        <termSec>
            <pron notation="AGP">myǫ̇lå̀r</pron>
            <gramGrp><gen>m</gen></gramGrp>
            <note>rapace diurne en général</note>
        </termSec>
    </langSec>

    <!-- multiple forms at one survey point-->
    <langSec xml:lang="fr-ALF-753">
        <termSec>
            <pron notation="AGP">ẹ̆klo</pron>
            <gramGrp><gen>f</gen></gramGrp>
        </termSec>
        <termSec>
            <pron notation="AGP">ọ̀kló</pron>
            <gramGrp><gen>f</gen></gramGrp>
        </termSec>
    </langSec>
</conceptEntry>
```

Fig. 2: TEI Dictionary specification

Fig. 3: TBX in TEI proposal

At the end of the Master Class, I hope to be able to decide which of the above data models – or which compromise thereof – best suits the dialectological data available in the *Atlas Linguistique de France*.

[9] https://www.iso.org/obp/ui/#iso:std:iso:16642:ed-1:v1:en

My name is Snežana Petrović and I would like to apply for the Lexical Data Masterclass in Berlin, December 4th till 8th.

I am a principal research fellow in the Department of Etymology at the Institute for the Serbian Language of SASA, currently working on two research projects supported by the Serbian Ministry of Education and Science:

- *Etymological study of the Serbian language and compiling the Etymological dictionary of the Serbian language*.
- *Interdisciplinary research of Serbian cultural and linguistic heritage. Creation of multimedial Internet portal "The Lexicon of Serbian Culture"*.

I am a principal investigator of the linguistic subproject (of the *Interdisciplinary* project) dedicated to digitization of dictionaries of Serbian dialectal and historical dictionaries. I am also a co-editor and co-author of the *Etymological Dictionary of the Serbian Language* (*Етимолошки речник српског језика*; Pilot issue 1998, First volume 2003, Second volume 2006, Third volume 2009) and an author of one monograph (*Turkish loan–words in the Serbian speech of Prizren*, Belgrade 2012), and more than 80 scientific articles.

I was also a head of two projects supported by the Serbian Ministry of Culture:

- Digitizing a paper slips word collection from Prizren by Dimitrije Čemerikić (http://www.prepis.org),
- Retrodigitizing Serbian dialectal and historical dictionaries (http://www.raskovnik.org).

I am a co-editor (with Toma Tasovac) of the Platform for digital edition and transcription of the Serbian manuscripts *Prepis* (http://www.prepis.org), co-author (with Toma Tasovac and Ana Tešić) of Digitized word collection from Prizren by Dimitrije Čemerikić (http://www.prepis.org/items/browse?collection=1), co-editor (with Toma Tasovac) of the Serbian dictionary Platform *Raskovnik* (http://www.raskovnik.org) and co-editor (with Toma Tasovac, Ana Tešić and Sonja Manojlović) of digital editios of two retrodigitized dictionaries *Rečnik kosovsko-metohiskog dijalekta* (*Dictionary of the Kosovo-Metohija Dialect*) by Gliša Elezović (http://raskovnik.org/recnici/GE.RKMD) and *Rečnik govora južne Srbije* (*Dictionary of* *the vernaculars of Southern Serbia*) by Momčilo Zlatanović (http://raskovnik.org/recnici/MZ.RGJS).

I am also a Member of the Etymological Committee (International Slavistic Committee) (2012–), the representative of the Institute for the Serbian Language in the DARIAH-RS, a part of the DARIAH-EU (2015–) and was a MC Member at the European COST Project IS1305 (European Network of e-Lexicography, ENeL) 2013–2017.

Since I have a certain level of experience in TEI based representation of lexical data I believe that this Masterclass can significantly contribute to improving both theoretical and practical aspects of my future work. The material I would like to use for the Lexical Data Masterclass is excerpted from *Rječnik iz književnih starina srpskih* (*Dictionary from the Serbian literary antiques*) by Đuro Daničić. It has three volumes, published in 1863 and 1864. The dictionary includes lexical material from a large number of older Serbian monuments, such as hagiographies, Letopisi srpski (The Serbian Chronicles) by J. Šafarik, Monumenta serbica by F. Miklošič, etc. Apart from giving semantic equivalents, in both Latin and Serbian, it provides various examples from the above-mentioned monuments in the form of syntagms or sentences. Daničić's dictionary is a monumental work of Serbian as well as Slavic lexicography, and one of the few existing dictionaries of the Old Serbian and Serbian Slavic language. This fact makes him a unique and important monument of Serbian philology and culture, but, at the same time, a very difficult and complex material for encoding in TEI.

We are planning to add this dictionary to the *Raskovnik* dictionary platform and I have already started working on the material with Ana Tešić. I strongly believe that the Masterclass could provide us both an opportunity to learn how to solve some of our problems, apply up-to-date approaches and exchange experience with other colleagues. In short – I'm sure It could make our online edition better. In the long term, however, both my colleague and me would like to expand our skills beyond text encoding: we would like to jearn the basics of Xpath so that we can perform more complex searches directly on our sources and we would like to experiment with automatic segmentation of etymological data from the *Etymological Dictionary of the Serbian* Language. The Berlin Masterclass is an ideal opportunity for us to acquire new knowledge so that we can share that knowledge with our other colleagues in Belgrade.

Sincerely yours,

Snežana Petrović

Alena Witzlack-Makarevich (University of Kiel)

Proposal

The scholarly background

Since Mai 2013 I am employed as an assistant professor (junior professor) for linguistic typology and language variation at the Institute for Scandinavian Studies, Frisian and General Linguistics (Department of General Linguistics) at the University of Kiel (Germany). My current research interests comprise two domains: linguistic typology and description and documentation of endangered languages. The second domain is more relevant to the present application. I studied general linguistic at the University of Leipzig (2001–2006). In 2006–2010 I worked as a research assistant in the project *Typological Variance in the Processing of Grammatical Relations* (University of Leipzig) and wrote my PhD thesis on *Typological variation in grammatical relations* (defended in January 2011) in the framework of this project. In 2011–2013 I worked as a postdoc in a follow-up project first at the University of Leipzig and since April 2012 at the University of Zurich.

My interest in lexicographic work relates to my second research focus on language documentation and language description. Already as a student assistant I participated in several field trips working on an undocumented variety of Richtersveld Nama (a Khoe-Kwadi/Central Khoisan language of South Africa). In 2009–2016 I carried out several field trips documenting the highly endangered language N|uu (Tuu/Southern Khoisan, South Africa). Among other topic, together with my colleagues I collected a lexicon of N|uu. This is a moribund language with a high degree of lexicon attrition, so that the nearly exhaustive lexicon comprised only 3000 words. At the moment this lexicon is being prepared for the publication.

In July 2015, I participated in the COST summer school *Standard Tools and Methods for Retrodigitising Dictionaries* with a project on D. Bleek's (1956) *A Bushman Dictionary.* This summer school provided me with the necessary theoretical and technical background and was pivotal for my participation in the application of another lexicographic project: In 2016 together me and a number of colleagues from the Makerere University (Kampala, Uganda) applied for a Volkswagen foundation project within the Initiative *Knowledge for Tomorrow – Cooperative Research Projects in Sub-Saharan Africa*. The project *A comprehensive bilingual talking Luruuli/Lunyara-English dictionary with a descriptive basic grammar for language revitalisation and enhancement of mother-tongue based education* (abbreviated in the following as the LLED Project) was funded starting from January 2017 (PI: Saudah Namyalo, https://www.isfas.uni-kiel.de/de/linguistik/forschung/projekte/LLED/luruuli_lunyara). As a member of the LLED project I am responsible for the management of the lexicographic database and preparation of the on-line version of the dictionary. Below I will briefly present the project and the type of the data at hand.

Project description

The term Luruuli/Lunyara (ISO 639-3: ruc) is used to refer to at least three distinct but closely related language varieties viz. Western Luruuli, Eastern Luruuli, and Lunyara. The language belongs to the Great Lakes Bantu (Narrow Bantu, Niger-Congo) group of languages and is mainly spoken in the Nakasongola and Kayunga districts of central Uganda. The number of speakers is difficult to determine, the number of people registered as ethnical Baruuli/Banyara is about 190,000 according to the 2014 census (Simons & Fennig 2017).

Uganda is a multi-ethnical and a multilingual country, with 56 recognized ethnic groups (Uganda Constitution 1995) and at least 43 different languages, mainly of the Niger-Congo and Nilo-Saharan language families (Simons & Fennig 2017). Although English and Swahili are the only official languages, regional languages like Luganda (or Ganda) serve as a major lingua franca. This multilingual setting, dominated by English and larger, regional contact languages, makes survival difficult for small languages, such as Luruuli/Lunyara. In addition, the Baruuli/Banyara have only been recognized as independent indigenous communities in 1995. Due to a number of historical events related to the colonial history of Uganda, they had been considered to belong to the dominant Baganda ethnicity for almost a century. During that time, they were strongly encouraged to abandon their language and culture and adopt the Baganda way of life instead. As a result, the usage of Luruuli/Lunyara was reduced almost entirely to the domestic setting, whereas Luganda and English were used in most other contexts. For this reason, Luruuli/Lunyara is considered to be a threatened language (Ethnologue, Simons & Fennig 2017).

Until a few years ago, there was no standard orthography for Luruuli/Lunyara and it still is mainly an oral language. First publications in Luruuli/Lunyara emerged a few years ago and at present the community works on publication of further texts in Luruuli/Lunyara. In addition, no language description exists to date (cf. http://glottolog.org/resource/languoid/id/ruul1235). In 1995 the Baruuli/Banyara were recognized as two independent ethnic groups with their own kingdoms, this led to an increased interest of the community members in the promotion and revitalization of their language, e.g. by making it the teaching language of primary schools and producing the first publications.

Luruuli/Lunyara-English dictionary

It is at this point that the LLED project takes up and aims to support the community's efforts in their endeavor to make Luruuli/Lunyara the language of primary education in their schools and other domains. Uganda's education policy provides the necessary legal basis for the education in the primary years 1–3 in the mother tongue provided that it has an approved orthography and literature. Appropriate teaching materials are also essential for successful education. The policy may sound promising, but it is difficult for communities, such as the Baruuli/Banyara to fulfill these requirements without any help from government institutions or NGOs and lack of funding. Therefore, the dictionary and grammar sketch compiled with the LLED Project will not only contribute to language documentation in general, but also to the revitalization and acceptance of Luruuli/Lunyara within the community, as well as to a better education for the Baruuli/Banyara children.

By the end of 2020 the LLED project promised to deliver the following items:

a) a lexical electronic database of approximately 10,000 entries (in Toolbox format)

b) 500 printed copies of the Luruuli/Lunyara-English dictionary of approximately 10,000 entries

c) an online talking dictionary of approximately 10,000 entries published as e.g. a contribution to the electronic open-access journal Dictionaria (http://home.uni-leipzig.de/dictionaryjournal/)

d) a mobile app of an adapted version of the dictionary

For practical reasons, at the moment the lexical data are collected and stored in the Toolbox format. Toolbox is a data management and analysis tool for field linguists (https://software.sil.org/toolbox/). Despite its many drawbacks, it is particularly popular in the filed of the documentation of endangered languages, as it is easy to use and allows for the simultaneous collection of lexical data, parsing and annotation/interlinearization of texts. In addition to using this format, the project regularly uses Python scripts to regularly convert the growing dictionary into the csv-format and use it with other applications.

Toolbox dictionary files present plain text files with individual fields starting with a tag. The LLED project uses the following major fields (followed by a number of other fields).

\lx	lexeme
\a	alternative forms
\ps	part of speech
\ge	English translation
\pl	plural form (for nouns)
\cite	citation form
\lxid	unique ID of the lexical item
\so	language (to track cases of code switching)

Further fields containing detailed grammatical information and examples will be added in the future. Specifically, we need to link individual lexical items to a database of example sentences (ideally with an audio recording). As is typical of the Bantu languages, Luruuli/ Lunyara words are extremely complex morphologically and can contain up to ten prefix slots and up to five suffix slots. The citation form regularly provided by the speakers contains several inflectional affixes in addition to the lexical stem and this discrepancy should be systematically dealt with in the dictionary. Nouns can belong to one of 20 genders/noun classes and the respective noun class is pivotal in e.g. determining the plural form of the noun. Both the noun class and the plural formation pattern should be captured in the dictionary. A verbal root allows for multiple morphological derivations (causative, reflexive, applicative) of varying productivity and semantic transparency, this present another challenge for the dictionary. Such structural considerations will determine the final structure of the lexical entries.

For the on-line publication of the dictionary we initially suggested to use the Dictionaria framework. This is a framework for online publication of peer-reviewed dictionaries of minor languages (http://home.uni-leipzig.de/dictionaryjournal/). The framework provides a web application for viewing and searching. The basis of this functionality comes from the relational database format of the dictionary: It can include up to four linked tables plus multimedia content (e.g. audio files). The Dictionaria project provides instructions for an easy transfer of dictionaries in Toolbox format into a relational database format, which allows a quick online publication within an established framework. However, as it is not guaranteed that the Dictionaria project will continue in the future and it is not yet clear whether it will provide the necessary functionality. For this reason, I am interested to considered other options for the online publication of the dictionary.

My tasks within the LLED project related to the online publication of the Luruuli/Lunyara include workflow and database management, data conversion from the Toolbox format to whatever format we will choose for the publication and sustainability management. In addition I will be responsible for including the grammatical information into the dictionary.

Many of the topics mentioned in the call for applications for the Lexical Data Masterclass. are of relevance to the present project and the timing is ideal, as many decision related to the organization of the entries and the format have not been ade yet. The relevant topics include dealing with morphology in a digital lexicon, managing examples in dictionaries, managing digital lexica as online resources, the adherence to the TEI guidelines and TEI based representations for lexical content, and actually most of the other suggested topics.

My name is Ana Tešić and I would like to apply and hopefully, take part in the Lexical Data Masterclass in Berlin, December 4th till 8th.

I hold a Ph.D. degree in Linguistics (Etymology) from the Faculty of Philology, University of Belgrade while at the same time holding the position of research assistant at Serbian Academy of Sciences and Arts' Institute for the Serbian Language (Etymological Department).

I was involved in a project aiming to create a database of Serbian cultural heritage that will be searchable on multiple levels and will include different segments of intangible cultural heritage called *Prepis* (http://prepis.org). This is a joint project of the Institute for Serbian Language and the Belgrade Center for Digital Humanities. It includes the lexicon manuscript collected by Dimitrije Čemerikić in the middle of the 20th century documenting the now almost vanished Serbian dialect from the historic city of Prizren. It is a searchable database, as all the scans are given titles according to the manuscript or, in cases where that wasn't possible, according to lexicographic methods. Some of the scans are transcribed, improving the searchability of the database. Data is enriched with synonyms and standardized word forms. The list of semantic fields and subfields is formed to allow easier thematic tagging of the collection.

Another project, currently undergoing, is the *Raskovnik* project (http://raskovnik.org). This is another infrastructure project run by the Institute for the Serbian Language of SASA and the Belgrade Center for Digital Humanities as part of DARIAH-RS, aiming to create an expanding and technologically advanced environment for the digitization of Serbian linguistic heritage. Another aim is to eventually develop the tools needed for the research of the Serbian language and culture. The idea is to create a dictionary platform for the digitization and interconnection of various Serbian dictionaries, to provide a comparable search of all the existing dictionaries in the database (by lemmas, grammar forms, definitions, examples, etc.), to encourage an even deeper research of the Serbian lexicon and finally, to develop different language tools for computational linguistics as well as for the needs of library systems, such as indexing, text annotation, etc. By doing so, it will hopefully contribute greatly to further popularization of the Serbian language and culture. There are currently three dictionaries on the platform – *Srpski rječnik* (*Serbian Dictionary*) by Vuk Stefanović Karadžić (both editions, from 1818 and 1852), *Rečnik kosovsko-metohiskog dijalekta* (*Dictionary of the Kosovo-Metohija Dialect*) by Gliša Elezović and *Rečnik govora južne Srbije* (*Dictionary of the vernaculars of Southern Serbia*) by Momčilo Zlatanović, with other dictionaries currently being prepared to be uploaded to the platform.

The material which would be used for this masterclass would be excerpted from one of the dictionaries that will eventually have its place on the platform – *Rječnik iz književnih starina srpskih* (*Dictionary from the Serbian literary antiques*) by Đura Daničić. It has three volumes, first two published in 1863 and the third one in 1864. The dictionary includes a lexis of a large number of older Serbian monuments, such as hagiographies, Letopisi srpski by Šafarik, Monumenta serbica by Miklošič, etc. Apart from giving word interpretations, in both Latin and Serbian, it provides various examples from the above-mentioned monuments in the form of syntagms or sentences. This dictionary is a monumental work of Serbian as well as Slavic lexicography and stands as a monument of Serbian philology and culture.

I would be working on this material alongside with my work colleague from the Institute for the Serbian Language, Snežana Petrović. Since we are both involved in the project I believe it is important for us both to participate actively together in the Master class in order to achieve later on the best possible results with our project.

I believe this masterclass would provide me with the opportunity to upgrade my theoretical knowledge as well as improve the practical experience I have so far, which would be of great value in my future work. My Institute would also benefit greatly from my training in this field, as we are currently in the process of introducing digital editing in our workflows with a focus on digitizing, transcribing and mutually linking various dialect dictionaries in different projects. Dialect and historical dictionaries require a great deal of careful scholarly annotations considering that they document non-standard and/or archaic vocabulary. I am convinced that TEI provides a flexible enough framework for making semantically rich editions that will make our scholarly work more productive.

I expect the Lexical Data Master class to be a true challenge, but one that I have a great interest in pursuing. I strongly believe this workshop would be very beneficial to me and I would be pleased to be given a chance to participate in it.

Once again I am grateful for considering my application and I look forward to a favorable reply.

Application for LexMC: Lexical Data Master Class

Franziska Diehr

October 19, 2017

In my profession as an information scientist I am engaged in the development of metadata schemes, data models and controlled vocabularies in the context of Digital Humanities projects. With a background in working on the documentation and management of cultural heritage objects and collections, I am experienced in dealing with data about artefacts, their historical background and provenance, involved processes, agents and places. For some time now, my interest in linguistic research questions emerged and I became curious about working with text corpora, especially methods and techniques for their preparation and analysis. My particular research interest is how to model and deal with uncertainties, vagueness and ambiguities.

1 SCHOLARLY EDUCATION AND BACKGROUND

During my studies of museology (B.A.) at the HTW Berlin from 2007 to 2010, I developed an interest in the documentation, organisation and management of knowledge. This was one of the reasons why I wanted to do my Master's in Information Science at the Humboldt University of Berlin (HU). I desired to know more about the theory of information and wanted to learn how to model, manage and retrieve information. In 2011 I started working as a student assistant at the Hermann von Helmholtz Center for Cultural Techniques (a central institute of HU). There, I worked for the BMBF funded project "Coordination Centre for Scientific University Collections in Germany"[1], which also served as an inspiration for my master thesis.[2] After

[1] http://wissenschaftliche-sammlungen.de

[2] Diehr, Franziska: Ontologisch basiertes Datenmodell für die Beschreibung wissenschaftlicher Sammlungen, Humboldt-Universität zu Berlin, Philosophische Fakultät I,/url013, http://dx.doi.org/10.18452/14209

receiving my master's degree in 2013, I had a one year fellowship from the Association of German Foundations, where I was engaged with the project "Exploration and Documentation of German Foundation Archives".[3]

2 Current Work and Special Interest in Lexical Data

Since 2014 I am working as a research assistant in the project "Text Database and Dictionary of Classic Mayan"[4], which aims to develop a corpus based dictionary. The project is a collaborative work of the Department of Anthropology of the Americas at the University of Bonn and the State and University Library Göttingen, where I am member of the Metadata and Data Conversion group. In my responsibility lies everything concerning the project's needs in modeling, creating and managing scholarly information. This includes the development of metadata schemes and data models. Until now, we collaboratively developed 1) a CIDOC CRM based metadata schema for the documentation of text carriers and their historical and scholarly context, including controlled vocabularies modeled in SKOS, 2) a sign catalogue for mayan script, which provides a new concept for identifying graphemes, enables the documentation of graph variants and a way to deal with multiple reading hypotheses; and 3) a project specific TEI schema for encoding the corpus of about 10,000 Maya inscriptions. For creating, managing and saving the project's data we use the VRE TextGrid.

Our aim is to compile a dictionary. Therefore, my particular interest is on how lexical data can be encoded and extracted from the corpus. Since I am no linguist, my knowledge about lexical data and the requirements for the development of a dictionary is limited. To improve my skills on this subject, I am especially interested in Master Classes' teaching topics "models for lexical content (onomasiological vs. semasiological)", "dealing with morphology in a digital lexicon" and "TEI based representations for lexical content". In using the Oxygen XML Editor I have some practice. I know how to deal with the TEI Guidelines and how to write a TEI specification with ODD. In using XPATH and XSLT I have no experience.

3 Presentation of the Project's Data

When working with a not yet fully deciphered script, a sign catalogue, that is used for identifying the glyphs used in a specific text, is an essential working instrument for the epigrapher. Since all published sign catalogues concerned with Maya hieroglyphs struggle with false and multiple sign classifications, we wanted to overcome those difficulties by designing a digital sign catalog for the project. The catalogue also works as a basis for the development of the corpus in TEI/XML. Because we have to deal with multiple reading hypotheses and a huge variety of graph variants, we decided not to encode phonemically transliterated text, but every glyph with a reference to the URI of the graph in the sign catalogue. What we gain is a machine-readable corpus, which consists of instances of the sign catalogue. A further step of data processing enriches the corpus data with the transliteration values given in the sign catalogue. The outcome is a human readable text, ready to be linguistically analysed. The

[3] http://stiftungsarchive.de/

[4] http://mayadictionary.de/

linguistic analyses of the text corpus form the basis for the dictionary. Further questions are how to combine the results from the linguistic analyses and the corpus data into a dictionary and how to develop a dictionary, which enables the reference to the original text passages and allows researchers to enter lexicographical notes and analyses.

Since the project has only just started to create the corpus, we cannot provide a huge dataset to be worked with in the Master Class. If a few TEI-coded inscriptions would be sufficient, we would gladly share them. All our project's data and outcomes are licensed under CC-BY-4.0, so it can be used in the context of DARIAH-teaching material.

The Lexical Data Master Class would be an excellent opportunity for me to broaden my knowledge in the field of text data, in particular to deepen my knowledge on lexical data and to improve the application of technologies of the XML family. Further the Mayan dictionary project would benefit a great deal, if I get the chance to discuss the data and requirements with the experts and attendees of the Master Class. It would help us to model the lexical data for the Mayan language in such a way that the researchers' requirements for a corpus-based dictionary can be met.

Proposal

Summary

Lexicographic representations of German (language) are based almost exclusively on written language. In the LeGeDe project (**L**exik des **ge**sprochenen **De**utsch) at the Institute for German language in Mannheim we are building a corpus-based resource of spoken German by benefiting from methods of corpus linguistics and conversation analysis. At the Lexical Data Master Class I would like to learn the basic skills used in modelling a resource of spoken language that could provide an access from a semasiologic (lexeme-oriented) and an onomasiologic (topic-oriented) perspective.

Scholarly Background

Since September 2016 I have been working as a research assistant at the Institute for German Language in Mannheim, Germany (IDS). Currently I am working on projects 'Lexicon of Spoken German (LeGeDe)' and '**O**nline-**W**ortschatz-**I**nformationssystem **D**eutsch (OWID)', both located in the Lexis department. My main tasks consist in developing methods for lexicographic analysis of a corpus of spoken language (Research and Teaching Corpus of Spoken German; FOLK) and working on modelling and on an internet-presentation of corpus-based dictionary of spoken German. Prior to IDS, I studied Slavonic languages and Literature (BA) and Multilingual Text Analysis (MA) at the University of Zurich. Shortly after finishing my MA, I have worked part-time on creating a Levelled Study Corpus of Russian (LeStCor) at the Institute of Slavonic Studies in Hamburg, and as an assistant at the URPP Language and Space in Zurich.

During my MA studies I particularly focused on creating, annotating and performing linguistic analysis of a textual corpora. I am familiar with the different methods of annotating language data (lemmatising, PoS-tagging, parsing, etc.). I have spent a significant part of my studies working with parallel corpora and machine translation. As a student assistant at the Institute of Computational Linguistics in Zurich, I have worked with different XML-technologies and database modelling. Thanks to my linguistic background in Slavonic languages, I learnt to analyse corpus-data both qualitatively and quantitatively.

Topic

The topic I am currently working on is concerned with the creation of a lexicographic resource that provides information about salient terms in everyday spoken German. The resource is intended to provide access to new and additional meanings of everyday terms, which are not described in other dictionaries that are built on examples taken from written corpora. The project's contribution is the creation of a prototype of a lexicographic resource that describes common practices and preferences in spoken German used in different types of conversational settings, ranging from private to institutional and public categories. We are particularly focusing on lexical preferences related to frequent terms in spoken language, such as verbs, which are used differently in spoken than in written communication, different kinds of interjections, particles and multiword expressions. We are aware of only one similar project on spoken Danish (Hansen & Hansen 2012), and another one being currently developed in Slovenian (Verdonik & Sepesy Maučec 2017).

Dataset

We base our analyses on the Research and Teaching Corpus of Spoken German (FOLK) that can be accessed through the Database for Spoken German (DGD). FOLK is the largest corpus of spoken German in interactional context (1.95 million tokens). For analysing the data for dictionary content, we extract corpus examples from the DGD and analyse it according to various lexicographic and pragmatic criteria. Accept analysing categories such as mental, perception and communication verbs as well as different particularly related to style and register, we are also exploring lexicographic representation for more challenging types of entries, such as fixed expressions (*passt schon, hör mal, müssen wir mal gucken*) and delexicalised verb forms such as *passt, stimmt* or *geht*, none of which have been extensively elaborated in German lexicographic tradition. One of the main challenges of this type of resource consists in modelling word nets providing an access from an onomasiologic perspective, which would include lexical preferences for expressing affiliance, indetermination, doubt, culpability, disapproval, rejection, etc. In addition, since our corpus material is provided with metadata about speakers and events, we are planning to integrate this information into the lexicographic descriptions as well. At the Lexical Data Master Class I am hoping to 1) learn more about modelling lexicographic entries and exploiting the possibilities of their digital form, and 2) develop a prototype for different types of dictionary entries that will be extended in further work on the lexical resource of spoken German.

References

Hansen, C. & Hansen, M. H. (2012). A Dictionary of Spoken Danish. In R. V. Fjeld & J. M. Torjusen (eds.) Proceedings of the 15th EURALEX International Congress. 7-11 August 2012. Oslo, Norway: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 929–935.

Verdonik, D. & Sepesy Maučec, M. (2017). A Speech Corpus as a Source of Lexical Information. In International Journal of Lexicography. 30 (2), pp. 143–166.

Proposal for the Lexical Data Masterclass

I graduated from Istanbul University, English Language Teaching department in 2002 and received my MA degree in 2006 from Yıldız Technical University. In my MA, I studied Teaching Turkish as a Foreign Language and pedagogical lexicography. My thesis focused on constructing a multilingual online dictionary for learners of Turkish. The structure of the multilingual electronic dictionary was based on a model I developed, called "Çokdilli Öğrenici Sözlüğü Modeli - Multilingual Learners Dictionary Model". This model was based on the background database work of an electronic dictionary created from scratch that semi-automatically matches words from different languages by using the data entered by lexicographers and already matched pairs. The system does not solely depend on matches done by the SQL server, instead these matches are proposed to the admins of each language. Therefore, the nature of this model is semi-automatic. This model proposes that multiple word matches from different languages could be achieved by using pairs from already matched pairs from two languages and extending it into other languages by way of creating networks between words and their matches in other languages. For this project, I used mainly SQL server based cron jobs to populate results for matching word pairs from different languages that are entered into the database. As individual dictionary items for each language are entered into the dictionary, in this case into the database, the system begins matching words with other languages that have already been matched in at least with one other language. In this way, a multilingual dictionary is developed out of monolingual dictionaries by time. This project may seem pretty basic as of now, but in 2006 this was a big step for creating a multilingual electronic dictionary in Turkish as there was neither a literature nor a completed work in practice back then on this specific subject. Recent advances in technology and in machine learning may help this model to become fully automatic rather than a semi-automatic nature, if it can be turned into a project in the coming years as long as the dictionary items are being clearly annotated and tagged. The problem of such a system in MLD (Multilingual Learners Dictionary) model is that machine matching does not come up with satisfying results, but it does provide some easiness for the lexicographers to work on. This factor is one of the limitations in my MA work and the second limitation would be to find the admins for multiple language pairs who are capable of approving the right matches in practice. I tried to build a website for this project using the theoretical concepts taken from my MA thesis but due to development and maintenance costs I was unable to put it into practice.

Recently, I am doing my PhD at Ankara University and am focusing on the typologies of dictionaries in general; how they are prepared and what type of items are necessary for each type of dictionary. The goal of my PhD study is to specify the components and items that are needed for a dictionary to become a dictionary and create an electronic platform that can help lexicographers to create any type of dictionary in whatever category it could reside in appropriately. These items constitute a general framework for any dictionary in a broader sense, if the lexicographer decides to compile a learners' dictionary then the framework would allow the lexicographer to select the minimum items that classifies the dictionary as a learners' dictionary. If the lexicographer tries to compile another type of dictionary, e.g. a terminology dictionary, then the framework would allow the lexicographer to select the necessary items that can classify such a dictionary as a terminological dictionary. The framework would also be very flexible as new multi-disciplined dictionaries may emerge in the market. Therefore, I am working on the components that are to be annotated and tagged in order to build such a framework. The first problem that I encounter now is that there is no unified typology for classifying dictionaries; some sources make their classifications according to the languages being used in the dictionary, like multilingual and monolingual dictionary, and some sources make their classifications according to the information the dictionary cover in the definition of an entry.

In order to achieve such a classification, I started to study dictionaries that can be considered “official” dictionaries of all major languages in the world. For example, in Turkish, I mainly focused on the one published by the Turkish Language Institution which acts as the official authority on the language (without any enforcement power) and contributes to linguistic research on Turkish. The institution is charged with publishing the official dictionary of Turkish, "Türkçe Sözlük (Turkish Dictionary)". There is an XML version of the dictionary. In order to use the XML file, I started to learn XSL and XQUERY all by myself in order to be able to utilize the information in the dictionary, like words ending with a certain suffix or their syntactic categories. In the XML file, the tags are written in Turkish and the file itself is constructed in a non-standard way which has to be corrected using the ISO standards, because it creates problems for extensible use of the file. To be able to surpass the difficulties I had, I tried to convert some part of the XML information by using XSL to dump the data into MySQL database where I felt more confident with. After trial and error learning, I was able to create an XSL style sheet for the XML dictionary file and prepared a MySQL database out of it. This task was not satisfying because I was able to convert some part of the data. Therefore, the XML file should be corrected to make full use of it.

I believe that this master class will provide me the ability to exchange ideas with other researchers who are interested in lexical data and machine-readable dictionaries, so that we can share ideas on how we are working on data sets with certain tools and how we overcome certain difficulties in specific languages, like Turkish, and make each other feel more comfortable using the necessary tools, computer languages and the standards to create, manage and disseminate lexical data. I hope to attend the masterclass so that I would be able to give feedback to other researchers who may be interested in studying Turkish lexical data.

You can find the supplementary files below (and their URLs):

1) XSL file that I used for converting the data from XML into SQL. I used underscore “_” as the delimiter when creating SQL tables in the import file (dictionary-xsl-delimiter_.xsl) https://www.dropbox.com/s/dl645yalwkxnlmo/dictionary-xsl-delimiter_.xsl?dl=0

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
   xmlns:xs="http://www.w3.org/2001/XMLSchema"
   exclude-result-prefixes="xs"
   version="2.0">
      <xsl:template match="/sozluk/kelime">
                        <xsl:for-each select="grup">
                           <xsl:value-of select="grup_ID"/>_<xsl:value-of
select="grup_bilgi"/>_<xsl:value-of select="grup_anlam"/>_<xsl:value-of
```

```
select="grup_atasozu_deyim_birlesikfiil"/>_<xsl:value-of select="grup_birlesiksoz"/>
                                </xsl:for-each><br />

        </xsl:template>
</xsl:stylesheet>
```

2) A representation of the data that I am working on (dictionary-xml-sample.xml)
https://www.dropbox.com/s/i2mhltzjphbj0gk/dictionary-xml-sample.xml?dl=0

```
<?xml version="1.0" encoding="windows-1254"?>
<sozluk>
      <kayit>
                  <kelime>ab</kelime>
                  <grup>
                              <grup_ID>ab</grup_ID>
                              <grup_bilgi>isim, eskimiş</grup_bilgi>
                              <grup_anlam>
                                          <anlam>Su.</anlam>
                              </grup_anlam>
                              <grup_birlesiksoz>
                                          <soz>abıhayat</soz>
                                          <soz>abıkevser</soz>
                                          <soz>abuhava</soz>
                              </grup_birlesiksoz>
                  </grup>
      </kayit>
</sozluk>
```

I have a Ph.D. in General Linguistics and work at the University of Helsinki with open-source finite-state morphological description of minority languages in the Giellatekno infrastructure, based at the Norwegian Arctic University, in Tromsø, Norway. In my work, involving mainly Uralic languages, I utilize hand-crafted xml dictionaries that are xsl-transformed for producing necessary code lines necessary in transducers[1], on the one hand, but whose multilingual glosses with semantic and syntactic notation can also be used in morphology-savvy online/click-in-text dictionaries[2], ICALL environments[3], and syntactic disambiguation. Since not all language learners and users are versed in xml editing[4], I have made efforts in bring synchronic editing via MediaWiki to the language community[5]. The MediaWiki environment also allows for semantic searches and multilingual access[6].

The languages I am working with have been researched in German, Russian, Finnish, Estonian, Hungarian, Swedish, Norwegian, French and more recently in English. The National Library of Finland has conducted a Kindred language digitation project between 2012 and 2017 with extensive open-source digital materials available for language research and revitalization stemming from the 1920s and 1930s[7]. This means that there is a wealth of materials and research results that can be applied to the multilingual facilitation of large number of Uralic languages. And it is my intension to expand on the dictionaries through intensified cooperation with the Finnish Language Bank in Helsinki, dictionary development at Giellatekno[8], FU-Lab[9] in Syktyvkar, Komi Republic, Russia, as well as Udmurt language research in Turku, Finland and Szeged, Hungary[10].

As one of the driving principles behind work on the Giellatekno infrastructure is minimizing the number of times things are written/coded, I have taken it upon myself to maximize the usability of our dictionaries. I also believe in making lemma, stem, inflection, derivation, semantic, audio and etymology data available for multiple reuse. This has been achieved to a great extent through use

[1] Erzya, Komi-Zyrian, Moksha, Hill Mari, Nenets, Olonets-Karelian, Skolt Sami ... e.g. https://victorio.uit.no/langtech/trunk/langs/sms/src/morphology/stems/

[2] Click-in-text dictionaries at Giellatekno provide access to minority language .html texts, e.g. Erzya and Moksha http://valks.oahpa.no/ , Skolt Sami http://saan.oahpa.no/ and also Permic languages http://kyv.oahpa.no/.

[3] ICALL follows the lead of work with North Sami http://oahpa.no/davvi see work overseen in Skolt Sami http://oahpa.no/nuorti

[4] Work with Komi-Zyrian xml-s can be seen in the Giellatekno infrastructure at https://victorio.uit.no/langtech/trunk/words/dicts/kpv2X/inc/

[5] Doctoral student Mika Hämäläinen has worked to facilitate synchronic editing of the Giellatekno xml materials in MediaWiki at CSC and the Finnish Nation Language bank, e.g. the Skolt Sami word for priest: https://sanat.csc.fi/wiki/Sms:papp with export of nouns at: http://sanat.csc.fi:8000/smsxml/xml_out/?language=sms&type=morph&file=N_sms2x.xml

[6] By adding a basic Finnish word list of adjectives, adverbs, nouns and verbs to the https://sanat.csc.fi infrastructure in Helsinki, we have been able to access semantic equivalents from languages with articles pointing to Finnish, e.g. the Finnish word *lepakko* 'bat (mamal)' can access semantic equivalents from 5 languages: https://sanat.csc.fi/w/index.php?title=Toiminnot%3AT%C3%A4nne+viittaavat+sivut&target=Fin%3Alepakko&namespace= (what links here)

[7] National Library of Finland Kindred Language digitation project harvest: https://fennougrica.kansalliskirjasto.fi/

[8] Ciprian Gerstenberger works with dictionary and corpora infrastructure at Giellatekno Tromsø, Norway: https://en.uit.no/ansatte/organisasjon/ansatte/person?p_document_id=77186&p_dimension_id=88149

[9] My work with Komi-Zyrian goes back to 1996 with a very simple dictionary Komi-Finnish-English, but has continued to the coordination or several separate languages. It has also taken me to collaboration with http://dict.fu-lab.ru/ in Syktyvkar.

[10] Sirkka Saarinen has conducted Udmurt-Finnish and Finnish-Udmurt projects; the materials are accessible at: https://victorio.uit.no/langtech/trunk/words/dicts/udmfin/README and ongoing work by István Kozmács in Szeged, Hungary: https://victorio.uit.no/langtech/trunk/words/dicts/udmhun/README

of TEI-derivable xml-structures developed in the Giellatekno infrastructure, extensively applied in the language documentation projects I have been involved in[11]. These data, when coordinated in parallel for several languages, provide resources for both science and layman knowledge, which is otherwise out of the reach of minority language development structures.

I realize that my collaboration in xml-based dictionaries for many of the Uralic literary languages may lack optimal consistency, hence a review of TEI-based representations will be beneficial. In addition finding and choosing examples representative of valency/government/semantic based divisions in the dictionary structures, as well as hints for utilizing fieldwork research dictionaries, will be of great use. I am continually looking for new ways to empower minority language communities and want to enhance their feeling of community through orchestrated dictionary development. For this reason, I am looking forward to participation in the master class because I think many of my questions have ready answers that I am not aware of.

The dataset I would like to enrich during the master class is the XML lexica for Skolt Sami[12] in the Giellatekno infrastructure. This lexical dataset consists of multiple XML files each of which corresponds to one part-of-speech and described in xsd[13]. Consistency will also play an important part, as there exist extended structures used in the mutual xsd for other languages, i.e. Erzya, Komi-Zyrian and Udmurt.

I am currently working with the merging of multiple translation languages into shared entries. The work flow here is important for automation work done by Mika Hämäläinen at the University of Helsinki with the merging of morphosyntactically and semantically defined meaning group subelements of the xml dictionaries. This is work in progress, and any work done with it in the master class context can be seen as beneficial and open-source available in the very same open-access Giellatekno repository.

[11] Open-source morphological projects Jack Rueter is involved in: Erzya, Skolt Sami, Komi-Zyrian, Livonian, Olonets-Karelian, Moksha, Hill Mari, Meadow Mari, Nenets, Udmurt, Votic, etc.

[12] Sanat dictionary platform in the Finnish CSC infrastructure: https://sanat.csc.fi/wiki/Sms/sokk and the tandum Norwegian Giellatekno infrastructure: https://victorio.uit.no/langtech/trunk/langs/sms/src/morphology/stems/ as well as https://victorio.uit.no/langtech/trunk/words/dicts/finsms/

[13] The xsd has been developed but for lack of input from others using dtd representation this needs development https://victorio.uit.no/langtech/trunk/giella-core/schemas/fiu-dict-schema.xsd

simon.gabay@unine.ch

Université de Neuchâtel
FLSH - 3.O.14
Espace Louis Agassiz 1
2000 Neuchâtel
Suisse

20 October 2017

Dear Madam, dear Sir,

After a bachelor in Paris IV-Sorbonne (France), a licence in St Andrews (Scotland), a master in Paris IV-Sorbonne (France) in French literature, I have written a PhD at the University of Amsterdam (the Netherlands) in latin philology about lexicographic issues (the vocabulary to designate the actor in medieval latin). I am now a postdoctoral fellow of the Université de Neuchâtel (Switzerland), where I work on a digital edition of Mme de Sévigné's autograph manuscripts (17th c. French letters).

The main goal of my postdoc is to adapt the editorial practices of mediaevalists to 17th c. French texts. Contemporary editions of the latter are indeed very problematic from a philological point of view: their spelling is massively regularised, no attention is given to manuscripts, and linguistic studies (*e.g.* glossaries) are never offered to the reader. I have therefore thought my edition of Madame de Sévigné as the opportunity to discuss and propose a new editorial protocol, dealing with transcription rules, the apparatus and the annexes (linguistic study, glossary, *etc*.)

My project, funded by the Fond National Suisse is currently in its third (and final) year. Extensive documentary research has been carried in libraries of the United States, France, England, Germany, Austria and Italy, but also on the private market to gather as many autograph manuscripts (or facsimiles) as possible. Letters are currently being transcribed and edited, and the bulk of the autograph correspondance should be available on time (October 2018).

I am now starting to think on the question of the annexes, and especially the most complex one: the glossary. There is indeed no scientific dictionary for 17th c. French: specialists usually use Richelet (1680), Furetière (1690) or the one written by the Académie (1694), which is problematic. To address the issue properly, I am working with the director of the *Französisches Etymologisches Wörterbuch* (FEW) in Nancy (prof. Yan Greub), with whom I will discuss a first draft of my work in two weeks.

The FEW is now part of the *Analyse et Traitement Informatique de la Langue Française* (ATILF) lab of the CNRS, and is therefore at the forefront of digital humanities — dictionaries created in this lab, when they are not digitally native (*e.g.* DMF), are carefully retroconverted (*e.g.* FEW). During a research trip in Nancy, I was lucky enough to meet some of the members of the IT staff specialised in annotating text with part-of-speech and lemmatisation, who accepted to help me with my data.

My project rely indeed massively on digital tools, and my long term goal is to create a web portal on 17th c. manuscripts, for which Madame de Sévigné's *Correspondance* is a test case. I have thought from the start this digital catalogue as a great opportunity to carry lexical research, since it would be one of the few opportunities to study words of the 17th c. in their manuscript context, and provide new data on spelling or graphical features.

In order to achieve this goal, I have followed many courses of digital humanities in many different countries (England, France, Germany), and plan to keep doing so for the following months. My critical edition of Madame de Sévigné is entirely encoded in XML-TEI, but in such a way that it could easily be expanded with new authors. The code will soon be inspected by a specialist of virtual libraries to assess its quality, before the publication of a beta version of the website in February.

However, linguistic, and especially lexicographic aspects, are still a problem relatively hard to tackle for me, and attending *LexMC: Lexical Data Masterclass* would be of a great value at a critical moment of my project. Annotation within the transcription through tokenisation, linked to a separate glossary is the option on which I am working now, but it can still be modified.

Unfortunately, my XML-TEI encoding is still very uncertain for the glossary, because of the complexity of the task. The idea is to offer more than a simple definition: a small commentary discussing the status of the word, its history and its uses, would therefore be provided. To do so, the glossary would merge definitions of 17th c. dictionaries (Richelet, Furetière, Académie, Trévoux), scientific historical dictionaries of French (*FEW*, *DMF*, etymological section of the *TLFi*) and other studies (on Sévigné for instance, cf. Fr. Nies's work on her vocabulary). The selection of the relevant informations, both in the source texts and in the dictionaries, to provide a correct description and definition of the word is difficult — especially since these informations should eventually be linked to the future digital TEI version of the *FEW*.

The analysis of orthographical variation is one of our priority, since, at the time of the *Querelle des anciens et des modernes* (Quarrel of the Ancients and the Moderns), spelling is political. The distribution of specific features in time and places (modernisms, archaisms, *etc*.), sometimes even throughout the life of a writer, will be treated with care, and (if possible) related to the glossary.

The idea of proposing a non-digital version of our work (*e.g.* pdf produced via XSLT from the TEI encoded document) would be of a great value. However, no commercial use of it would be envisaged, since I am attached to the principle of free of charge publications. All the material I would bring to Berlin will therefore be available to everyone (a draft with a few dozens of entries heavily commented and the transcriptions), if you accept my application.

The importance I give to sharing knowledge goes however beyond free publications. Since I am organising a winter school in Neuchâtel (February 2018), during which I will address the question of XML-TEI and lexical data, everything I would learn would be transmitted to the persons attending the course — and especially my colleagues of the *Glossaire des patois de la Suisse romande* (*Glossary of the patois of Romandy*), with whom I am working closely in Neuchâtel.

In the meantime, I remain at your full disposition should you have any questions or need any further information.

Best regards,

Simon Gabay

Applicant: Paraskevi Savvidou
Position: PhD Student in Linguistics, National and Kapodistrian University of Athens, Greece.
Email: psavvidou@phil.uoa.gr

Academic and research background

I hold a BA Hons in Greek Philology with major in Linguistics from the Aristotle University of Thessaloniki (Greece) as well as a Master of Science in Theoretical and Applied Linguistics from the National and Kapodistrian University of Athens (Greece). I am currently a last year PhD Student in the Department of Linguistics of National and Kapodistrian University of Athens. In my PhD thesis I examine the evaluative morphology in Modern Greek from a corpus-based perspective. The aim of my dissertation is to contribute to the description of the evaluative morphology of Modern Greek and the wider discussion of the 'peculiar' character of evaluative morphology cross-linguistically, as well as to the understanding of corpus linguistics theory, by offering a new view on the neglected interface between corpus linguistics and word formation study. Part of my research has been published and presented in several conferences; the most recent ones are the *9th International Conference on Corpus Linguistics* (9th CILC, June 2017, Paris) and the *13th International Conference on Greek Linguistics* (13th ICGL, September 2017, London). In the context of the latter (13th ICGL, London), I also co-organized a workshop on Evaluative Morphology in Greek. Moreover, recently (June 2017) I participated in ENel Training School which held in Waterford, Ireland and I presented my research in its poster session. I have also worked as a researcher in various projects in greek Universities (Aristotle University of Thessaloniki for the years 2012 and 2013, National and Kapodistrian University of Athens for the years 2013-2015) and Research Institutions (Academy of Athens, 2010, 2012), mainly in the fields of corpus linguistics and lexicography. More specifically, I was a member of the research team that designed and compiled the first Diachronic Corpus of Greek of the 20th Century (Greek Corpus 20) as well as a member of the team that compiled one of the first learner corpora of Modern Greek (ESKEIMATH, GLC, Greek Learner Corpus). Additionally, I have voluntarily contributed to the enrichment of the Corpus of Greek Texts (CGT), the most representative general corpus of Modern Greek, with a sub-corpus of approximately 30,000 words from the genre of short messages (SMS).

General description of an ongoing (meta)lexicographic project

My main research interests lie within the fields of Word formation study, Corpus Linguistics and Lexicography. In my thesis, as well as in my wider research, I explore the ways in which corpus linguistics can revise the metalexicography and lexicography of word formation study, in a way parallel to that in which corpus development changed our view on lexicography in general, with emphasis on phraseology. I attempt a historical overview of corpus linguistics theory and practice in order to demonstrate that word formation is a neglected area both as an object of theoretical study and in its lexicographical description. I explore the reasons for this lack of interest, I offer an explanation of it, I try to suggest the ways to overcome the limitations of current research on corpus-based morphology and I associate them with the fundamental principles of corpus approach; both the reasons of the limitations of corpus morphology as well as the ways to overcome them are found at these principles. My aim is to develop a methodology which will transfer all the benefits of the 'phraseological' approach of corpus linguistics to word formation study and its lexicographical description. The first results of that part of my research have been presented in various conferences in Greece and abroad. The lexicographic part will be developed further in the post-doc period of my research and will be the extension of the ideas and assumptions of my doctoral thesis; it would have as an outcome a comprehensive methodology for applying corpus techniques in word formation study, in a particularly more extensive way than in existing research. Further outcomes will be the lexicographic description of word formation processes of Modern Greek, with emphasis on derivation and compounding, as well as an extensive review of the lexicographic works in Modern Greek, both traditional and modern, which will indicate the neglected points of the description and the impact that the proposed corpus methodology could have on them.

Relevance of the proposed project to the MasterClass

As a part of the above project, I will attempt the digitization of a part of the 'Great Dictionary of the Greek Language', one of the most important dictionaries of the Greek Language, which was directed by Dimitrios Dimitrakos and it was published in 9 volumes from 1936 to 1950. My aim will be the exploration of the contribution of lexical data of that type to the investigation of word formation processes of Greek as well as of their (meta)lexicographical description. Therefore, I will concentrate on the parts of the dictionary that represents the word formation rules of greek, with emphasis on derivation and compounding. My potential participation in the Master Class will contribute significantly to this project, as it will give me the opportunity to have a guidance and feedback from experts in the field on the technical aspects of my project.

More specifically, during the master class, I will have digitized all the lemmas of 'Dimitrakos Dictionary' that have as a first element the morpheme "theo" (θεο-), which comes from the noun "theos" (θεός) which in Greek means "God" and as a compound part may denote the meaning "god" but it may also express intensification (e.g. from a "Dimitrakos" entry: "theopsilos" (*θεό-* 'god' and *ψηλός* 'tall'), which means very tall. The entries of the compound lemmas with "theo-" in the above lexicon extend to fourteen (two-columned) pages. My dataset in the Master Class will be a txt file which will contain the plain text of these fourteen pages after a digitization which will begin with the application of an OCR on the scanned pages and will be completed with the required corrections of the text in order to avoid the conversion errors. During the Master Class, I wish to encode the above part of the dictionary in a TEI schema. This schema will be customized in order to be suitable for that specific part of a dictionary which represents the word formation rules of a language. More specifically, it will aim to be suitable both for the lexical entry of the sub-lexical unit "theo-" as well as for the entries of its compounds. The information of each lexicographical entry will be marked up with a set of recommendations which will be link the compounds with the lemma of "theo-" as well as with the other compounds of "theo-", by indicating their similarities in meaning etc. Therefore, the aim of my participation in the Master Class will be the creation of a pilot XML-TEI conformant source which will be based on a sample of lexical data. In the future, this pilot schema will be applied to all the entries that represent the word formation of greek language (compounding, derivation).

This project could have a significant contribution both to the field of word formation study of Modern Greek as well as to a critical review of the history of the greek (meta)lexicography of word formation. More specifically, the proposed source of lexical data could be analyzed both qualitatively and quantitatively in order to extract information about the description of compounding and derivation; our research questions will deal with issues like the potential asymmetries in the lexicographic description regarding the kind of the meaning of the elements (descriptive or evaluative) as well as regarding its grammatical status (compound, derivative), its productivity etc. These findings will be compared with the results of Modern Greek dictionaries, in order to explore the history of word formation and its lexicographic description diachronically. Moreover, the findings will be compared with the results of a corpus-based analysis of the same elements. The expected outcomes of this project will include the proposal of a methodology for a dictionary of word formation of Modern Greek, which will be benefited by the extensive review of the existing lexicographic works. Moreover, the project will be a significant progress in the field of available lexical sources for the greek language. It is remarkable that the digitization of the important dictionaries of Modern Greek is limited to a simple scan of their content.

My participation in the master class will offer me the opportunity to attend lectures from experts and to be trained in practical issues regards the process of digitising in XML TEI a historical dictionary.I will also have the opportunity to discuss my ideas for the further development of my project and to have feedback regarding the technical aspects of the peculiarity of the word formation (meta)lexicographic descriptions.

Author Index