ELEXIS Report

ELEXIS Transnational Research Visit Grant

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

Travel Grant: Call 1

Hosting institution: Instituut voor de Nederlandse Taal

Period of stay: February 4-8th, 2019

Researcher: Margarida Ramos

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

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

Project title: Knowledge Organization and Terminology: application to Cork

ELEXIS Report

ELEXIS Transnational Research Visit Grant

Introduction

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

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

Research goals

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

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

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

Overview

Day #1

Meeting the Head of the INT, prof. dr. Frieda Steurs

Scheduling of the meetings for the week

Acknowledging the team and departments

Day #2

Meeting subject: *Phraseology & Word combination*, with Colman, lic. Lut

NedTerm, a terminological resource, with Kinable, dr. Dirk

Day #3

Oral presentation of my PhD project methodology: *Knowledge Organization and Terminology: application to Cork*

Meeting subject: *Corpora & metadata editing*, with Depuydt, lic. Katrien creativity along with its counterpart, the labour-intensive data analysis tasks. It is this merge of creativity with scientific work that motivates my interest on contacting lexicographers and computational linguists, in short connection with informatics, i.e. a multidisciplinary collaborative team.

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

Brainstorming sessions

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

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

Day #4

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

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

Day #5

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

Farewell

(i) *Terminological records in the digital era*

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

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

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

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

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

Tools & technology, what else for e-lexicography?

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

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

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

NedTerm, a terminological resource

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

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

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

Conclusion

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

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

Acknowledgements

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

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