elexis european lexicographic infrastructure

PAN IJP Institute of Polish Language Polish Academy of Sciences

ELEXIS TRANSNATIONAL RESEARCH VISIT GRANT FINAL REPORT

Grant holder: Dorota Mika, PhD

Affiliation: Institute of Polish Language, Polish Academy of Sciences (Poland)

Hosting institution: Instituut voor de Nederlandse Taal (Netherlands)

Period: 30.05.2022-3.06.2022

Project Title: Integration of lexicographic data: the diachronic plane

Research objectives

The main problem currently confronting linguists using lexical resources is the significant fragmentation of information. Inconsistent resources fragment our knowledge of the language. Integrating data would make it possible to give a complete description of the language from the earliest times to the present day.

I am a researcher at the Institute of Polish Language, Polish Academy of Sciences (IJP PAN) which is the leading centre of lexicographic research in Poland. The collection of IJP PAN includes scholarly dictionaries of the Polish language, based on linguistic material collected by several generations of researchers – historical, dialectal, onomastic and present Polish dictionaries. This collection is highly heterogeneous, apart from electronic dictionaries (completed and still ongoing), it includes printed dictionaries, supplements, word card catalogues and corpora.

IJP PAN has collected a huge amount of lexicographic data. As a result, many challenges have arisen, such as how to store this data effectively and how to integrate the separate resources. Together with a team of IJP PAN employees, I am working on a research project called Dariah.lab: „Digital Research Infrastructure for Arts and Humanities" (POIR.04.02.00-00-D006/20; https://lab.dariah.pl/), where we are using existing and original solutions developed for the digitization, reconciliation and integration of lexicographic data.

The research objective is to develop and adapt methods for processing printed dictionaries (turning printed dictionaries into annotated digital versions), and for integration of lexicographic data. IJP PAN resources, provide a challenge to existing processing and integration methods. Firstly, the nature of the data is varied (highly structured dictionaries of the Polish language and extensive corpora). What is more, textual data is expected to be integrated with resources currently available in the form of images: the word card catalogues, there are thousands of paper and partially digitised cards, store fragments of historical or dialectal uses of the language.

Research Visit

Thanks to the e-Lexis project of visiting grants, I had the opportunity to visit the Instituut voor de Nederlandse Taal (INT) in Netherlands. INT studies all aspects of the Dutch language, including its vocabulary, grammar and linguistic variations. This institute has a long lexicographical tradition and a rich collection of lexicographical resources.

On the first day of my stay, there was a seminar in which I introduced the INT team to the tasks we are carrying out within the Dariah.lab project, related to processing printed dictionaries into a structured digital versions, enriching dictionaries, and integrating lexical resources. My host at INT, Katrien Depuydt, introduced me to the idea of a data-based integrated lexicographic infrastructure, that makes it possible to describe a language across the centuries and in all its complexity. INT is a leading center for e-lexicography. Projects for the integration, coordination, and stimulating the scientific description of language represent a high level of advancement.

Integrating lexical resources

A centralized model of language data management has been developed and implemented at INT to stimulate work on historical and contemporary resources. The INT collects data on grammar, terminology, neologisms, and dialects of the language. The different layers of language description form modules that together provide a complete overview of the language in its dynamics and complexity. The prepared workflow and the way the modules are organized enable the generation of multiple links between the resources.

The model applied at INT assumes a centralised management of resources. Data from dictionaries and corpora power a central lexicographic database. Particular elements of the entries build the layers of the diachronic module, that can be easily implemented in various projects. INT also provides users with an integrated dictionary portal, which makes it possible to search in all of these dictionaries. A wide set of linking facilities allows users to move freely between resources.

Linking existing resources: the historical dictionary portal

Four historical dictionaries, collected together, describe the Dutch language from about 500 to 1976: Oudnederlands Woordenboek (ONW, Dictionary of Old Dutch, 500-1200); Vroegmiddelnederlands Woordenboek (VMNW, Dictionary of Early Middle Dutch, 1200-1300); Middelnederlandsch Woordenboek (MNW, Dictionary of Middle Dutch, 1250-1550); Woordenboek der Nederlandsche Taal (WNT, Dictionary of the Dutch Language, 1500-1976). They provide the core material for the historical lexicographic infrastructure developed at the INT. The four dictionaries have been converted to a standardize TEI encoding and interlinked at the lemma level.

Modular approach to development of lexical resources

The GiGaNT lexicon has two main modules: GiGaNT Hilex – the historical lexicon component and GiGaNT Molex – modern lexicon component.

Work on modern Dutch is increasingly carried out in a centralized way. The *Algemeen Nederlands Woordenboek* (ANW) is a dictionary of contemporary Dutch, with advanced search functions. The ANW infrastructure connects to the project describing the newest Dutch vocabulary – neologisms. These two together are linked to the MoLex central module.

Conclusion

I came to INT to gain knowledge for processing and integrating lexicographic data from the diachronic perspective. The main research objective is to prepare the concept of integrating data from historical dictionaries created at the IJP PAN. I wanted to find the answers to the questions of how to integrate data from several dictionaries in an automatic way, and how to create the search for links at the level of headwords, word meanings, and inter-word relations to show the language from a diachronic perspective.

In Dariah.lab we are involved in complex lexicographic data processing – starting with OCR, moving on to post-correction and ending with the dictionary segmentation phase. Information in dictionaries is often abbreviated and highly condensed. We recognize the text using OCR tools (Tesseract and AbbyFine Reader). Tesseract, while giving quite good recognition results, does not preserve the typography of the text, which provides a lot of semantic information. During the visit, I was introduced to INT projects for which recognition of text and its structure has been a challenge, e.g. in the Couranten Corpus comprises the seventeenth-century Dutch newspapers (https://couranten.ivdnt.org/). Tools for automatic detection of regions, lines and words as Tesseract, Ocropy are useful in digitisation work on printed resources. Grobid and Transkribus can be used for layout analysis. The latter can be used both to correct recognized text and provide a better starting point for the semi-automatic structuring of dictionary entries.

Thanks to the e-Lexis project, I could see how a central language data management system was implemented here. INT's works on digitising printed dictionaries and integrating lexical resources are very advanced. The experience gained from this visit is a great help in the project currently ongoing. All ideas and suggestions will be discussed with the project team. The key to integrating dispersed resources is:

1. defining a consistent data model:
   a) identifying the modules (historical, modern, dialectal lexicon components);
   b) describing relevant relations between modules.
2. implementing an identifier system for electronic resources (lemmata, superlemmata, definitions, citations, and sources).
3. the construction of modules for the different layers of language (module for historical lexis, module for modern language, module for dialects), which are first linked within a module, then it is possible to create links also between these modules.

The works carried out on integrating the dispersed resources of IJP PAN are conducted within the Dariah.lab research project. As a result of integration, a unique database of lexicographic resources will be created. Access to the database will be possible via a modern and easy-to-use WWW interface.

Acknowledgements

I would like to express my sincere appreciation to all at INT, to Katrien Depuydt for her wonderful efforts as my host, for sharing with me her vision of modern e-lexicography and for showing me how this vision is being implemented at INT; to Carole Tiberius for preparing this visit and for all the valuable organizational tips; to Thomas Haga and Roland de Bonth for information about historical dictionaries at INT; to Vivien Waszink and Boukje Verheij for introductions to modern language projects; to Katrien and Jesse de Does for valuable advice on converting printed dictionaries into digital versions and for information on available tools; to Veronique de Tier for a presentation on dialect projects and introducing working methods; to Frank Landsbergen for a

presentation on grammar portals; to Dirk Kinable for a presentation on a terminology integration project. I would like to thank the INT team for their valuable support and for sharing their great experience with me.