Scientific Report of Short Term Scientific Mission

COST STSM Reference Number: COST-STSM-IS1305-010916-078225

Period: 01-09-2016 to 30-09-2016

Duration: 22 working days

COST Action: IS1305

STSM type: Regular (from Germany to the Netherlands)

STSM Title: Giving an overview of retro-digitised dictionaries and supporting the development of guidelines for retro-digitised dictionaries

Guest/STSM applicant: Ursula Schultze, Trier Center for Digital Humanities (TCDH)/ University of Trier

Host: Katrien Depuydt, Instituut voor de Nederlandse Taal (INT), katrien.depuydt@ivdnt.org

1. Purpose of the STSM

The aim of the STSM was to examine which dictionaries from the overview presented by Dr. Anne Dykstra (Fryske Akademy) during the COST EneL Meeting in Barcelona in March/ April 2016 were retro-digitised, and to develop a scheme to describe these dictionaries. Furthermore, numerous dictionary examples were collected which show how different projects realized their data modelling. The results contribute to the realisation of two goals of WG2: 1) creation of an overview of retro-digitised dictionaries - with the subtask of developing a scheme of categories describing retro-digitsed dictionaries; and 2) development of guidelines for the retro-digitisation process. It was possible to combine the STSM with my master thesis in which I developed a scheme to describe and order retro-digitised dictionaries

2. Description of the work carried out during the STSM

2.1 Preparation in Trier

With the scientific advice from Dr. Vera Hildenbrandt (TCDH) I decided to focus on retro-digistised dictionaries. After this I defined which retro-digitised dictionaries would be considered. Criteria for selection were: 1) the dictionaries were conceptualized as print dictionaries and have an printed equivalent and 2) they are fully accessible online so that the user is able to use the online version of such a dictionary without consulting the printed version.

A first guide to prepare the overview and the corpus for my master thesis was the inventory of European scholarly dictionaries. This overview was initially worked out by Gerbrich de Jong (Fryske Akademy) during her STSM at the TCDH in Trier in September 2014 which was supervised by Dr. Vera Hildenbrandt (TCDH) and Dr. Anne Dykstra (Fryske Akademy). I supported Gerbrich's work. Since Gerbrich's STSM, the inventory has undergone a continous improvement and expansion process. The recent version was presented by Dr. Anne Dykstra[1] (Fryske Akademy) at the COST EneL Meeting in Barcelona in March/April 2016.

[1] I would like to thank Dr. Anne Dykstra for providing me the access to this inventory.

A first step was to investigate which of the 259 registered dictionaries in this overview are retro-digitised. I then tried to determine which methods were used to transfer the printed version into a digital format and to discover which markup language was used to tag the dictionary data. This often was a challenge because in many cases there is no information on this on the project homepages.

To be able to give an accurate description of these retro-digitised dictionaries, I designed a first prototype of features and feature values which I presented Dr. Vera Hildenbrandt (TCDH) before my STSM at the INT started. I decided to chose three perspectives: a metalexicographical aspect, a technical aspect, and a media-specific aspect. I chose the metalexicographical aspect because it is important to know what the scope of a dictionary is, what a dictionary is about and how the entries are constructed. The technical aspect was chosen because, regarding retro-digitising dictionaries, it is important to find out how the transfer from the printed version into a digital format was realized and how the data were modelled. This information can be used to gain an overview of how the digitisation process was realised. This overview can be used again to recognize a development of a standardised workflow if there are tendencies of using very similar or very different methods. The media-specific aspect was choosen to show how and to which degree new possibilities on the internet are used to present the dictionary content and to present the dictionary access and the dictionary content more userfriendly.

With regard to my master thesis in German linguistics, I decided to split the dictionaries into a core corpus and a supplementary corpus which are relevant for my work. The core corpus includes German dictionaries, the supplementary corpus includes all dictionaries which deal with other European languages and are listed in the inventory. Against the backdrop that the inventory included only 6 dictionaries which are German dictionaries I decided to enlarge my core corpus to 20 dictionaries. I added 13 further German dictionaries which are important in the research discourse in the field of German lexicography and which have a printed equivalent and are online available. Furthermore, I decided to expand the core corpus to show the variety of scholarly retro-digitised dictionaries for the German language.

2.2 Work at the INT

In a first meeting with my supervisor Katrien Depuydt (INT) I told her about the challenge to determine which methods were used for the retro-digitsing process and which markup was used to model the dictionary data of the listed dictionaries in the inventory. We decided that I would write an e-mail to the concerning projects with a request for information on the methods and the markup language. Many staff members of the respective projects were very cooperative and supported my work with very detailed answers. Sometimes I got further suggestions of important scholarly dictionaries which had been retro-digitised but were not in my overview. When they fit my criteria, I added them to the list. At this point it is important to note that the overview can be expanded by any number of dictionaries. There are numerous digital libraries[2] which have dictionaries as part of their mass digitisation projects. These resources could be used to expand the overview. But providing a complete overview of these digitised dictionaries was beyond the scope of my STSM and my master thesis.

Furthermore, I worked on features and feature values to describe the dictionaries. Katrien Depuydt (INT) supported this work with scientific advice and helpful comments. The features and feature values with the specific definitions are attached to this scientific report in a Word file. After the final definition of the features and feature values, I collected all relevant dictionaries in an Excel sheet. The Excel sheet includes 109 dictionaries: 20 dictionaries in the core corpus and 88 dictionaries in the supplementary corpus.

[2] For German dictionaries there is http://www.digitale-sammlungen.de/, for Dutch dictionaries there is http://dbnl.org/, for dictionaries of all European languages there is https://archive.org/ and https://books.google.com/. There are some more resources available which cannot all be shown here.

We decided to interlink the entries with the database from the European dictionary portal built by PGD Michael Měchura (Fiontar, Dublin City University), which was a result of his STSM at the INT in April 2015[3]. The portal is based on the corrected and expanded overview which I also used. By interlinking the entries of my overview with the entries of the portal, a potential elaboration of the portal with information from my overview of retro-digitised dictionaries is enabled.
Furthermore, I marked the dictionaries which I further added to the overview.
I filled in all the features and feature values for the dictionaries that met my criteria. I coloured in the fields I was not able to work out. I was not able to fill in the metalexicographical aspect from dictionaries which deal with a language I cannot speak fluently and understand clearly in written form. I was not able to fill in some technical aspects when there were no information given about the used methods for the retro-digitising process and/ or the markup language to tag the dictionary data.
I decided further to expand the overview with retro-digitised dictionaries which did not meet my choosen admission criteria but are part of the corrected and expanded overview and are important to the research discourse in the field of lexicography and should not be rejected. These dictionaries are not part of my analysis and my evaluation.
If I was not sure about wether a dictionary was retrodigitised or I could not get further information, I marked them in red.

The analysis of the inventory is attached to the scientific report in an Excel sheet.
It would be advisable to present the overview to the MC-members so that they can review or complete the information for the dictionaries of their own language

2.3 Results and additional values

In a further step I transfered the analysed dictionaries into a new list. In this list I calculated the absolute and relative values of the filled-in feature values.
Following results can be shown:

Technical aspect
Feature: Digitisation

- 88% of the dictionaries are machine readable dictionaries
- 12% of the dictionaries are machine visible dictionaries

Feature: Method of acquisition

- 23% of the dictionaries were available in text documents, 9% thereof completely in text documents and 12% partly in text documents combined with other methods (e.g. scanning and/ or keying)
- 16% of the dictionaries were scanned
- 25% of the dictionaries were scanned and further processed with an OCR software
- 33% of the dictionaries were keyed
- 0% using HTR to transform the printed version into a digital format
- for 21% of the dictionaries there is no information available about the method of acquisition

Feature: Markup language, data modelling

- 19% of the dictionaries have no content related markup
- 30% of the dictionaries were tagged in XML
- 21% of the dictionaries were tagged in XML/ TEI
- 11% of the dictionaries were tagged with other markup languages

[3] In this context I want to express my gratitude to Dr. Bob Boelhouwer (INT) for establishing the contact to PGD Michael Měchura (Fiontar, Dublin City University), and to PGD Michael Měchura (Fiontar, Dublin City University) for providing me access to his data from the database.

- for 18% of the dictionaries, no information are available about the markup language or the data modelling

Feature: Presentation

- 40% of the dictionaries are designed as a faithful online representation of the printed version
- 49% of the dictionaries are designed as a new online version
- 32% of the dictionaries are designed as a pictorial representation of the printed version

Media-specific aspect

Feature: Kind of retro-digitised dictionary

- 49% of the dictionaries are not digitally expanded
- 51% of the dictionaries are digitally expanded

Feature: Multimediality

- 87% of the dictionaries include text to support the presentation of the content
- 8% of the dictionaries include text and pictures to support the presentation of the content
- 4% of the dictionaries include text and sound-files to support the presentation of the content
- 1% of the dictionaries include text and pictures and sound-files to support the presentation of the content

Feature: Search strategies

- in 16% of the dictionaries no search is possible
- in 41% of the dictionaries the full text is searchable
- in 72% of the dictionaries a search for lemmas is possible
- in 28% of the dictionaries an extended search is possible

Feature: Hypertextuality

- 37% of the dictionaries have elements in the articles with hyptertext without information processing
- 46% of the dictionaries have elements in the articles with hyptertext with information processing
- 44% of the dictionaries have no elements in the articles with hyptertext[4]

Analysis of the filled in features show some tendencies in the retro-digitisation process: most of the retro-digitised dictionaries in the inventory are machine readable. The methods of acquisiton show that half of the projects have used OCR or keying to transfer the printed version into a digital format. Half of the dictionary data have been tagged with XML or XML/ TEI. Only a minor part has used other markup languages to model their dictionary data. Furthermore, 41% of the projects offer a full text search in the retro-digitised dictionaries. 72% provide a search for lemmas, and 28% present an extended search. Only in 16% of the dictionaries no search is possible.

The Excel file has been converted into an Access database[5]. This database presents an additional value for WG2, for researchers working with digitising printed dictionaries, and for scholarly institutions which plan to retro-digitise dictionaries. The database allows to search the overview with individualized search queries. The database is further a support for my master thesis. With the database I can combine different search queries which can help to arrange the dictionaries and to develop a system to categorize the dictionaries. The database is attached to the scientific report. It is necessary to have Microsoft Access for using the database.

2.4 Personal value of the STSM for my scholarly education

[4] The given relative values are based on the calculations of the 11/07/2016.

[5] Thanks a lot to Dr. Jesse de Does (INT) for supporting this working step.

For me as an early stage researcher and for my scholarly education, the STSM was a very enriching experience. During the preparation of the STSM and during my stay at the INT in Leiden I could learn from experienced researchers how to improve my skills in precise expression of circumstances and issues. I learned to discuss the issues and results of my research project on a high level. I expanded my knowledge in academic writing and formulating scientific issues in English. Furthermore, I got in touch with the logic of a database and building a database what is a very advantageous skill for me. Moreover, I had the chance to build and strengthen my personal network.

2.5 Intended / Future usage of my work

WG2 organizes the workshop »Toward Best Practice Guidelines for Encoding Legacy Dictionaries« in conjunction with DARIAH-EU (Digital Research Infrastructure for the Arts and Humanities) and PARTHENOS (a H2020-funded project »Pooling Activities, Resources and Tools for Heritage e-Research Networking Optimization and Synergies«) in November 17-19, 2016 in Berlin. For this workshop I collected various dictionary entries which will be used as examples and working material for tagging dictionary data. Furthermore, some of the examples will be uploaded to the Repository from WG2.
My working results from the STSM are used in my master thesis to develop a system to categorize the dictionaries.