Scientific Report of Short Term Scientific Mission

COST STSM Reference Number: COST-STSM-IS1305-310815-059476

Period: 31-08-2015 to 11-09-2015

Duration: 10 working days

COST Action: IS1305

STSM Type: Regular (from Germany to Poland)

STSM Title: Lexicographic hybrids: a European perspective

Guest/STSM Applicant: Nathalie Mederake, Göttingen Academy of Sciences and Humanities

Host: Krzysztof Nowak, IJP-PAN Krakow, krzysztofn@ijp-pan.krakow.pl

1. Purpose of the STSM

Vocabularies of European languages feature words with comparable forms and similar meanings. Most of these words have a common Graeco-Latin origin and are discussed under terms like "Europaeism", denoting words with Eurolatin heritage or interlexemes (cf. Keipert 2010: 636). Historical dictionaries provide information within this Eurolinguistic scope as they document such words as borrowings or loanwords of Greek or Latin origin. Yet the lexical transfer can never be a one-time procedure. In fact, the import of signifiers and the development of the signified is and has been an ongoing process.

The aim of this STSM was to figure out means of lexicographic documentation of this kind of Eurolinguistic phenomenon and to look for information in historical dictionaries that could be of use in this context. Pursuing such a task calls for a word-information-system located somewhere between a dictionary and a web portal. Such a system can indicate lexicographic hybridity as well as it may require genuine dictionary components and other sources contributing to the exploration of the given phenomenon. In that respect, the work carried out by Bruno Bon and Krzysztof Nowak, who created WikiLexicographica (Bon/Nowak 2013), was considered a feasible solution. Even more so as the said wiki is oriented towards academic (though not diachronical) dictionaries with an historic outline. During my STSM at IJP PAN in Krakow I was able to learn about an already existing model that uses MediaWiki and SemanticMediaWiki to align Medieval Latin dictionaries as well as make proper use of it for the purpose of my self-imposed task. Also, questions like mine are objectives of the task group 'Metalexicography' in Working Group 4 of the COST action ENeL, and we took advantage of my visit to discuss further cross-working group sessions.

2. Description of the work carried out during the STSM

a) Choosing appropriate entries from a set of historical dictionaries: examples chosen from WNT, OED, [1]DWB, [2]DWB, TLF and SJP served as use cases for my STSM. Due to the purpose of my task I wanted the entries to meet different criteria such as a comparable signifier in at least three languages. The selected entries of historical dictionaries should also provide data covering information on etymology, the first occurrence of a lemma in the respective language including data

of the source texts used for citation, and a corresponding sense and/or domain in which the first attestation of the lemma occurred. With these dictionaries I was able to find examples mostly provided as open sources or including sources that were provided by one of the institutes (Göttingen Academy of Sciences and Humanities or IJP PAN). I respected the issue of copyrights in so far as I worked only with pointers provided in scholarly historical dictionaries such as OED (first attestation, possible meaning etc.). In this context, my objective was not to copy-paste entries of such dictionaries.

b) Getting familiar with the syntax of Semantic MediaWiki: The programming process was closely supervised by and discussed thoroughly with K. Nowak. The task was to implement a selection of articles and define search queries, e.g. related to a lemma's etymological information, in order to display the lemma on a Eurolinguistic level. We either worked with matching templates already existing for WikiLexicographica or created new ones to fulfill the purpose of my task. To name a few: we used templates to provide general overviews as summaries, and full article structures, templates for source indication to extract spatiotemporal information that can be displayed on maps and timelines, etymological data, etc. We also used a reference system linking our information to other historical dictionaries or European dictionary portals; here we incorporated information provided by dictionaryportal.eu as well. In order to maintain the wiki, we made use of inline queries and concept pages to add a hierarchy or expose additional data to a potential user.

c) Peeks around the corner: I was able to gain insights into various dictionary projects and discuss their prospects for digitalization, namely the Atlas Linguarum Europae (introduced by P. Dębowiak), which is seen as a forerunner for Eurolinguistic research questions. Further points of discussion were digitalization processes (Lexicon Mediae et Infimae Latinitatis Polonorum; introduced by K. Nowak), the potential of historical lexicographical portal options (medialatinitas.eu; introduced by K. Nowak and B. Bon), and electronic writing systems (The great dictionary of the Polish language; introduced by M. Biesaga) of projects that exist under the roof of IJP PAN.

3. Description of the main results

3.1 Approach

I justified my initial approach via signifiers (and in each case a comparable signified) because of ongoing discussions in the field of Eurolinguistics (see above) but also by means of feasibility. During the STSM K. Nowak and I came to the conclusion that an approach making use only of the signified would also be possible and could be worked on with the results of another task group such as A. Villalva's "Vocabulary of Emotions". We would like to explore this idea on one of our next meetings.

3.2 Microstructural aspects in the WikiLexicographica

An article or entry in WikiLexikographica contains essential lexicographic information divided in different tabs. The basic view summarizes basic information such as part of speech, abbreviated sense definitions, a timeline and a map of first occurrences. We deliberately worked with different kinds of visualizations showing where and when an attestation appeared for the first time. In doing so, we encountered the problem with inconsistent bibliographical data that dictionaries provide, if they provide it at all (e.g. [1]DWB). We should also be aware of the fact that when assigning date and place to an attestation, it is usually the date and place of print. Therefore, especially geographical data may not correspond with an author's actual geographical setting and thus provide misleading

information. For example: Kant's life and activities took place in Konigsberg, but his collected works were printed in Berlin (cf. figure 1). Furthermore, we tried to code text genres based on extra-linguistic criteria. To this end, we grouped types of text in diachronic text prototypes based on the approach used in "The Helsinki Corpus". Subsequent tabs provide more detailed information on etymology (with an embedded query on words of the same origin) and an abbreviated version of every original article. For means of maintaining the wiki, we created concept pages and used them, for example, for different naming of genders (e.g. Femininum = féminin(e) = vrouwelijk) to make information comparable.

3.3 Etymology

In respect of a shared etymology or comparable etymological strings, we created summarizing information boxes. Here, we discussed a kind of two-case scenario:

a) A lemma exists because of some underlying Graeco-Latin vocabulary inherited by different European languages (which is the point of view followed by Eurolinguistics). Here we encountered different problems: should we take different language periods into account which serve as a source for lexical borrowings, or not? How can we distinct lemmas that originate from classical Latin (e.g. Lat. *functio*) or vulgar Latin (e.g. Lat. *bestia*)? How can indirect and direct borrowing processes be handled? For example, Dutch WNT provides us with the information that nl. *beest* originates from anc. fr. *beste,* but German [2]DWB tells us that med. nl. *beeste* is relevant for dt. *biest*. Of course, we could only deal with the information we were given by the dictionary entries: if a dictionary did not provide detailed information on a lemma's origin or on a donor language, we did not assign it to the article. Nonetheless, our aim was to test how such connections, that is borrowings, could be made as transparent as possible (cf. figure 2).

b) A second sample scenario describes the use of a lemma in several European languages being currently coined (nowadays mostly discussed under the topic of Anglicism etc.). The challenge presented itself in terms of programming as follows: the source of an Europaeism is not a part of the underlying (Graeco-Latin) vocabulary, to which we assigned the main namespace[1], but the headword and its adjunct article serves as a reference for other articles in the WikiLexicographica. In order to display this kind of information as clearly as possible, we came up with certain conditions for an etymology template (cf. figure 3).

3.4 Further research questions

At the end of the STSM we were able to draft further and more general research questions to our work. We hope such questions will lead to other research visits or can be discussed in a different scientific framework: what are the patterns of lexical borrowing on a European level? From which languages travel words in which age, domains or text genres? And how could we use a natural language processing approach to answer such queries? What are specific requirements for software and data? And how should they be presented in electronic lexicography?

[1] The main namespace has been reserved for so-called "super-entries", i.e., entries of a kind of unified dictionary which can serve as an index for all headwords (cf. Bon/Nowak 2013: 411). For my approach a super-entry page for the headword *diaeta*, had certain spatiotemporal information about the word in question that has been retrieved by means of the embedded queries. This information is presented in the form of timelines and maps of the attested word occurrences which have been extracted from respective dictionary entries (cf. figure 4).

4. Intended publications and meetings resulting from the STSM

The work described here will possibly lead to one or two publications. Results will also be presented on the next ENeL meeting in Barcelona (March/April 2016). We would also like to propose a cross-working-group discussion there (regarding WG1 and WG 4) with working examples from our STSM. As a further outcome I will also be able to attend a joint meeting of the French-German research project regarding alignment of academic Latin dictionaries which will be held in Paris in October 2015.

5. Other comments and outlook

Research results from other workgroups (e.g. WG 1 dictionaryportal.eu) and COST actions (IS1005 Medioevo Europeo) were made use of during this STSM. It proved that such results are necessary and valuable to deepen European scientific relations as well as frame further research questions. The STSM also allowed a metalexicographical approach applicable to other task groups within Working Group 4, which could be a starting point for further and fruitful discussion.

6. References

Bon, B. / Nowak, K. (2013): Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki. Proceedings of eLex 2013, 407-420.

Habermann, M. (1999): Latein – „Muttersprache Europas“. Zum Einfluss des Lateinischen auf den Wortschatz europäischer Sprachen. In: DU 51:3, 25-37.

Haß, U. (2010): In search of the European dimension of lexicography. Plenary paper, held at the fifth International Conference for Historical Lexicography and Lexicology, Oxford, 16-18 June 2010.

Keipert, H. (2010) Die lexikalischen Europäismen auf lateinisch-griechischer Grundlage. In: Hinrichs, U. (Ed.), Handbuch der Eurolinguistik. Wiesbaden, 635-660.

Munske, H. H. / Kirkness, A. (eds.) (1996): Eurolatein. Das griechische und lateinische Erbe in den europäischen Sprachen. Tübingen.

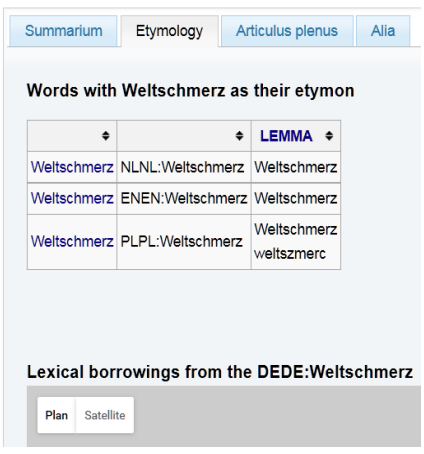
Simpson, J. (2004): 'Will the Oxford English Dictionary be more 'European' after its first comprehensive revision since its first edition in 1884-1928?'. In: Misceánea: A journal of English and American studies 29, 59-74.

7. Appendix

Figure 1: Visualising bibliographical data

Figure 2: Working with etymological information (scenario a)

Figure 3: Working with etymological information (scenario b)




```
[[Art etym::{{{1}}} {{{2}}}| {{{1}}} {{{2}}} ]]
{{#set_internal:Etymologia
|Etym string={{{1}}}{{{2}}}
|Etym lang={{{1}}}
|Etym source={{{2}}}
|Etym link=
{{#if:
 {{#ask:[[{{{1}}}]]
 |?Std dict
 |mainlabel=-
 }}
|{{trim|
 {{#ask:[[{{{1}}}]]
 |?Std dict=
 |mainlabel=-}}:{{ucfirst: {{{2}}} }} }}
|}}
|Etym link as a page=[{{#if:
 {{#ask:[[{{{1}}}]]
 |?Std dict
 |mainlabel=-
 }}
|{{trim|
 {{#ask:[[{{{1}}}]]
 |?Std dict=
 |mainlabel=-}}:{{ucfirst: {{{2}}} }} }}
|{{ucfirst: {{{2}}} }} }}]
|IsEtymOfLemme={{NAMESPACE}}:{{PAGENAME}}
}}
```

Figure 4: Example of a super entry