CDNL CENTER FOR DANSK NYLATIN
CENTRE FOR DANISH NEO-LATIN

DET KONGELIGE DANSKE
Videnskabernes Selskab
THE ROYAL DANISH ACADEMY OF SCIENCES AND LETTERS

Historical Lexicography in a Digital Age

Conference held at the Royal Danish Academy of Sciences
H.C. Andersens Boulevard 35, 1553 Copenhagen C
17-18 November 2022

Thursday 17 November	
15-15.15	Welcome by Marianne Pade, Aarhus University
	LINKING RESOURCES 1 Chair: Johann Ramminger
15.15-15.45	Marco Passarotti, Università Cattolica del Sacro Cuore, Milan: **The *LiLa* Knowledge Base. Linked Data Interoperability between Lexical and Textual Resources for Latin**
15.45-16.45	Lene Schøsler, University of Copenhagen: **Illustration of the Use of Lemmatization Strategies for the Analysis of a French Text from 1673** Gilles Souvay, ATILF/ Université de Lorraine: **Lemmatization Strategies on French of the 17th Century**
Coffee	
	LINKING RESOURCES 2 Chair: Eva Skafte Jensen
17.15-18.15	Achim Stein, University of Stuttgart & Carola Trips, University of Mannheim: **When corpus data may not suffice: The role of historical dictionaries in diachronic linguistics**
18.15-18.45	Thomas Troelsgård, Society for Danish Language and Literature: **A Meta Dictionary of Danish**

Friday 18 November	
COMPUTATIONAL SEMANTICS Chair: Marita Akhøj Nielsen	
10-10.30	Katrien Depuydt, Instituut voor de Nederlandse Taal (INT Dutch Language Institute): **Towards a Diachronic Semantic Lexicon of Dutch (DiaMaNT)**
10.30-11	Sanni Nimb, Society for Danish Language and Literature: **The linking of senses in two monolingual Danish dictionaries, the historical *Ordbog over det danske Sprog* and the modern *Den Danske Ordbog*: Results and Perspectives**
Coffee	
11.30-12.30	Marc Alexander & Fraser Dallachy, University of Glasgow: **Semantic Tagging with the *Historical Thesaurus of English***
Lunch	
REUSE OF HISTORICAL DICTIONARIES Chair: Lene Schøsler	
13.30-14	Peter Zeeberg, Society for Danish Language and Literature: **Renaessancesprog.dk – a Latin-Danish lexicographical website**
14-14.30	Marita Akhøj Nielsen, Society for Danish Language and Literature: **Publishing a dictionary with 300 years' delay. The retrodigitization of Matthias Moths' Danish dictionary**
14.30-15	Eva Skafte Jensen, The Danish Language Council: **Old dictionaries as a means to study historical grammar**
Coffee	
NEW STRATEGIES FOR EXISTING PROJECTS Chair**:** Marianne Pade	
15.30-16	Roberta Marchionni, Thesaurus linguae Latinae: **To make the *Thesaurus linguae Latinae* even more successful – Lexicographers and IT experts together at work**
16.30-17	Johann Ramminger, Thesaurus linguae Latinae: ***Neulateinische Wortliste*: Origins and Perspectives**
17-17.30	Simon Skovgaard Boeck, Society for Danish Language and Literature: ***Dictionary of Old Danish* – status and future possibilities**
Conclusions by Giancarlo Abbamonte, Università degli Studi di Napoli Federico II *Reception*	

The conference is funded by the Royal Danish Academy of Sciences and Letters *and the* Centre for Danish Neo-Latin

Abstracts:

Marc Alexander & Fraser Dallachy, University of Glasgow

Semantic Tagging with the *Historical Thesaurus of English*. The Historical Thesaurus Semantic Tagger (HTST) is the first semantic annotation software to account for the date of the input text when disambiguating possible meanings of a word. It uses the University of Glasgow's *Historical Thesaurus of English* (1965–), whose primary "parent" is the data of the *Oxford English Dictionary*, and in addition to the tagger the SAMUELS project produced two large-scale diachronic text collections: the *Hansard Corpus* (British parliamentary speech, 1803-2005) and *Semantic EEBO* (early print books from the Early English Books Online collection, 15th-18th Centuries). This talk will discuss lessons learned from the project, including the need to identify a medium-grained set of semantic categories for end users as well as for data aggregation, the variation in tagging accuracy depending on text date, and which improvements could be made to future iterations of the software. It will also give an update on ongoing work on the semantic composition of the *Hansard Corpus* and *Semantic EEBO*.

Simon Skovgaard Boeck, Society for Danish Language and Literature

***Dictionary of Old Danish* – status and future possibilities.** In recent years, work on the *Dictionary of Old Danish* has been suspended. This hiatus has offered an opportunity to scrutinize the principles behind DOD and to clarify ideas for future projects. The many recently digitized historical dictionaries provide inspiration for lexicographical developments of DOD, but some caveats are necessary. In particular, this paper focuses on possible connections between digital dictionaries and text editions and on interrelations between dictionaries.

Katrien Depuydt, Instituut voor de Nederlandse Taal (INT *Dutch Language Institute*)

Towards a Diachronic Semantic Lexicon of Dutch (DiaMaNT). In my talk I will discuss the ongoing development of a computational semantic lexicon of Dutch (DiaMaNT) at the INT. This lexicon is part of a larger lexicographical infrastructure for historical Dutch consisting of lexical data and corpora that we have been developing since 2005. The core of the infrastructure is formed by the four scholarly historical dictionaries of Dutch: the W*oordenboek der Nederlandsche Taal* , the *Middelnederlandsch Woordenboek* , the *Vroegmiddelnederlands Woordenboek* and the *Oudnederlands Woordenboek*, together covering Dutch language from ca. 500 – 1976. The main purpose of DiaMaNT is to enhance text accessibility and to foster research in the development of concepts through time. We will explain how the lexicon is linked to the other components of the lexicographical infrastructure and how we have developed its content so far. A first version of DiaMaNT was released in the fall of 2019 and can be found at diamant.ivdnt.org.

Eva Skafte Jensen, The Danish Language Council

Old dictionaries as a means to study historical grammar. This talk on historical lexicography will represent a user's perspective. Old dictionaries, of course, contain an abundance of words. They also to varying degrees contain grammatical information. The latter is very useful to the historical linguist who is interested in changes in the grammatical systems. In my talk, I will demonstrate how I use a number of historical dictionaries when studying changes in the grammar of Danish.

Roberta Marchionni, Thesaurus linguae Latinae

To make the *Thesaurus linguae Latinae* even more successful: Lexicographers and IT experts together at work. The Thesaurus linguae Latinae has been exploiting new technologies since 2003: first with the electronic version on CD Rom and then online (De Gruyter); more recently, with an open access version available on its Homepage. Since about a year, a team composed of lexicographers and IT experts has been working to transmit the acquired results in an even more efficient and wide-ranging way, also safeguarding the enormous amount of data produced during the work on an article.

The aim of this talk is to show the two main 'creations' and their advantages: 1) a 'Beleg-stelleneditor', a system that organizes the material and the most disparate acquisitions, creating 'Forschungsdaten' reusable by the authors of other articles; 2) an edition program for the article itself in XML language, which allows the user to draw from the Thesaurus much more information than has been possible until now and in a much faster way.

Sanni Nimb, Society for Danish Language and Literature

The linking of senses in two monolingual Danish dictionaries, the historical *Ordbog over det danske Sprog* and the modern *Den Danske Ordbog*: Results and Perspectives. The talk deals with research carried out in the project ELEXIS concerning the automatic linking of senses of identical lemmas in two monolingual Danish dictionaries, the historical *Ordbog over det danske Sprog* and the modern *Den Danske Ordbog* (DDO). Initially, the senses of 500 polysemous lemmas were manually linked across the two dictionaries before the dataset was used as a gold standard in automatic linking experiments. I discuss the results, but also take a closer look into the cases where there is no match between the two dictionaries. This information is relevant not only in the linking of other historical Danish dictionaries, but also in relation to studies of sense development in modern Danish. Finally, I will demonstrate how we automatically inherit information from a modern Danish thesaurus, *Den Danske Begrebsordbog*, when we link historical dictionaries to the DDO senses because this is already linked with data from the thesaurus.

Marita Akhøj Nielsen, Society for Danish Language and Literature

Publishing a dictionary with 300 years' delay. The retrodigitization of Matthias Moths' Danish dictionary. Denmark's first comprehensive mother tongue dictionary was compiled by Matthias **Moth** (1649-1719). He worked on the project from 1698 until his death, and his surviving material includes 62 folio manuscripts. It has been used by Danish lexicographers since the middle of the 18th century, but was otherwise known only to very few Nordic philologists. With the retrodigitization and online publication of Moth's final edition in 2013-2015, the user group

changed radically. In recent years, Moth's work has been used both as a source and as an aid in numerous historical investigations, but also – by the wider public – as an entertaining collection of funny and quirky words. The transplant from a small group of researchers to the general public, however, raises the question of the responsibility of historical lexicographers for completely anachronistic uses of the handed down substance.

Marco Passarotti, Università Cattolica del Sacro Cuore, Milan

The *LiLa Knowledge Base*. Linked Data Interoperability betwesen Lexical and Textual Resources for Latin. In my talk, I will present the structure of the LiLa Knowledge Base, i.e. a collection of multifarious linguistic resources for Latin described with the same vocabulary of knowledge description and interlinked according to the principles of the so-called Linked Data paradigm. Following its highly lexically based nature, the core of the LiLa Knowledge Base consists of a large collection of Latin lemmas, serving as the backbone to achieve interoperability between the resources, by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. After detailing the architecture supporting LiLa, the talk will particularly focus on how we approach the challenges raised by harmonizing different strategies of lemmatization that can be found in linguistic resources for Latin. The talk will also present a number of examples of queries run on the resources currently linked to LiLa.

Johann Ramminger, Thesaurus linguae Latinae

***Neulateinische Wortliste*: Origins and Perspectives**. The *Neulateinische Wortliste* (NLW) is a dictionary of Early Modern Latin from Petrarch to 1700. In continuous development since the 1990s, it has been freely accessible on-line since 2003. It has focused on 'new' words (i. e. words not or sparsely attested in Classical Latin), and on words with new meanings. With over 20000 entries it is by now not only the largest reference work for the Latin of the period in question, but also a unique resource for studies in word formation and language change in general. Entries are mostly short. The scholarly community has contributed lexical material and, more significantly, some entries (*humanitas*, *academia*) which contain thorough analyses of the semantics throughout the period covered by the NLW. The paper will highlight some of the research potential offered by the NLW and discuss possibilities for future expansion, especially the integration with other on-line resources.

Lene Schøsler, University of Copenhagen & Gilles Souvay, ATILF/ Université de Lorraine

Illustration of the Use of Lemmatization Strategies for the Analysis of a French Text from 1673 (Lene Schøsler). The second paper presents an illustration of the use of the tools for the analysis of a French text, an original ego-presentation written by the Danish princess Leonora Christina in 1673. After the transcription of the ms, the text was lemmatized in collaboration with Gilles Souvay. The work steps include the analysis proposed by the tool, disambiguation by the researcher (including separation or grouping of units), revision of labels, verification or correction of the transcription in the event of doubts raised by the analysis, and systematic control of lemmas for verification. Typical problems were: Unresolved words, ambiguous or unknown forms, words with an incorrect lemma, proper nouns. Certain words must be grouped to form multi-lemmas. In

other cases, there is the need of separation of merged units: not only words not separated by an author, but also the merging of articles, such as *au, aux, du, des …*

Lemmatization Strategies on French of the 17th Century (Gilles Souvay). The first paper provides a presentation of lemmatization strategies dealing with an inflected text language (French of the 17th century), the main problem being the graphic and morphological variability of a vernacular language from this period. The LGeRM lemmatization platform was used to process the text and make it available to the community. It is built around a lemmatizer capable of processing the old states of French (language not standardized or in the process of standardization). The lemmatizer relies on a lexicon (i.e. a list of lemmatized words from corpora) and on rules of graphic and morphological variations (based on general knowledge and on information from corpora). The platform makes it possible to visualize, disambiguate and possibly correct the results of the lemmatization. A consultation of the final text is possible from predefined interfaces allowing continuous browsing of the text and interactive access to both words and lemmas.

Achim Stein, University of Stuttgart & Carola Trips, University of Mannheim

When corpus data may not suffice: The role of historical dictionaries in diachronic linguistics. In this talk we will address three foci of the conference: the description of language use by means of NLP tools, lemmatization strategies, and retro-digitization of lexicographical resources. The introduction briefly presents our research background and our motivation. Our investigations mainly revolve around verb argument structure in a diachronic and acquisitional perspective and are carried out on the basis of historical annotated text corpora for both English and French. Since verb argument structure is a phenomenon at the interface between syntax and the lexicon, the shortcomings of textual databases regarding word-level annotation, especially lemmatization, are a serious impediment. In the main part of our talk, we will show how existing historical dictionaries (the *Middle English Dictionary*, MED, and the *Altfranzösisches Wörterbuch* by Tobler & Lommatzsch, TL) can be linked with syntactic treebanks of historical corpora and thus contribute to solving this issue. For both resources and languages, we will discuss lemmatization as a prerequisite for historical lexical-semantic analysis and possible links to current NLP resources like semantic networks. For the MED, a web-based tool kit (the *BASICS Toolkit*) assists users, among other functions, in building verb classes for the Middle English period, or in performing a reverse lookup on the MED entries. For the TL, we will show how retro-digitization, even though incomplete, provides not only the information required for lemmatization but also the connections with *WordNet* and other semantic resources.

Thomas Troelsgård, Society for Danish Language and Literature

A Meta Dictionary of Danish. The Society for Danish Language and Literature (DSL) produces Den Danske Ordbog (The Danish Dictionary), a comprehensive online dictionary of modern Danish. The society also hosts a series of retro-digitised dictionaries of older Danish. The Meta Dictionary is an attempt to interlink these dictionaries at lemma level, thus creating a central in-house resource that could be used for e.g. semi-automatic lemmatisation of older texts or for facilitating lookup functions across dictionaries describing Danish of a certain period.

In my presentation, I will describe some of the challenges encountered when linking headwords from quite heterogeneous resources. For instance, a loan-word like *fersken* ('peach') occurs in forms like *pfirsken, persike, pærsik, persick,* and *pfersing*, which, although etymologically related, are rather different in terms of form. Another example is variation in derivation, cf. cases like *forførersk*/*forførerisk, mindelse*/*mindsel,* or *gudfrygtig*/*gudsfrygtig/gudfrygtelig*. Finally, I will demonstrate how the Meta Dictionary is used in a website lookup-functionality.

Peter Zeeberg, Society for Danish Language and Literature

Renaessancesprog.dk: a Latin-Danish lexicographical website. The site Renaessancesprog.dk, which was published in 2009, combines a corpus of Latin and Danish renaissance texts with a web-edition of seven Latin-Danish or Danish-Latin vocabularies from the 16th and the beginning of the 17th centuries. The edition of the vocabularies represents an attempt to make seven structurally different dictionaries searchable as one. It encompasses alphabetical as well as systematic vocabularies and both Latin-Danish and Danish-Latin vocabularies.