An automatically generated Danish Renaissance Dictionary

Building a period dictionary by reducing and merging relevant existing dictionary resources

Mette-Marie Møller Svendsen, Nicolai Hartvig Sørensen, Thomas Troelsgård
Society of Danish Language and Literature

The Danish Renaissance Dictionary is part of the project "Music and Language in Danish Reformation Hymns" (2018-21) of which the main goal is to present a digital edition of a series of Danish hymn books from the Lutheran Reformation (officially implemented in 1536). Historically this was a time where many political, social, economical and religious changes were taking place. The Danish language was also experiencing a transitional period, which is of particular interest to this project. Luther's German translation of The New Testament in 1522 motivated the Danish King Christian II to initiate work on a Danish translation (Nielsen 2017). From then on more and more texts in Danish as well as hymns and services in Danish followed. The Reformation encouraged the use of the Danish language, compared to the use of Latin and Low German, and provided a significant boost to the expansion of the vocabulary.

Texts as well as music are digitised and made searchable, and a series of dictionaries relevant to the period, which have only partially been digitised, will be made accessible. Furthermore, for the key texts of the project, the text reader will include an integrated dictionary function that looks up the selected word and presents a generated basic entry – a sort of "sense keywords" – extracted from the project's dictionary sources. Links will also be provided to the full entries of the dictionaries where the selected word is found. The project is funded jointly by the Carlsberg Foundation and the Velux Foundation, and the work takes place at the Society for Danish Language and Literature (DSL).

In this presentation, we will describe how we perform the linking of the dictionaries, and the present stage of our work on processing and presenting data for the "keyword entries" in the Renaissance dictionary.

Dictionary linking

The available dictionary data is highly heterogeneous, as the project dictionaries comprise 12 dictionaries and vocabularies from the time of the Reformation as well as five more recent dictionaries and vocabularies (19th, 20th and 21st century) describing the language of the period. The material each dictionary and vocabulary offers differs greatly in scope and size, but generally (and naturally) the more recent dictionaries are much larger than the early source material. Furthermore, the uneven and sometimes even patchy levels of details and accuracy in the markup of the older dictionaries and vocabularies is another obstacle in the process.

For the linking task, the heterogeneity becomes apparent through a rich variation in spelling across the resources, as well as the provision of part-of-speech information (which is often absent in the older dictionaries), and the choice of base form of the headword. Hence, some resources list verbs in the infinitive, while others use the present tense. For these reasons it became evident that the linking could not be performed in a fully automated way.

The linking is done within a "meta dictionary" that is continuously in development, and which ideally aims at linking all Danish dictionary resources at DSL at entry level. The same meta dictionary is used for linking two modern Danish dictionaries in connection with our tasks in work package WP2 of the ELEXIS project. The work package is centered on dictionary linking across languages and achieving compatible formats for the ensuing meta dictionaries.

The linking of each resource is done in three passes:

1. If a source entry matches a target entry in the meta dictionary, having matching headwords and matching part-of-speech, and neither of them has homographic headwords (of the same part-of-speech) in their respective dictionary, the linking is considered safe and is completed automatically.
2. If one or more possible targets can be found in the meta dictionary, selecting the correct target is done manually using a custom-designed tool called the "Konnektor".
3. If no match can be found, the headword in question is established as a new lemma in the meta dictionary.

For the actual linking we use the tool "Konnektor". Its input is an XML file with a series of sets, each holding the entry to be linked and one or more possible target "meta entries" in the meta dictionary. The targets are organised by prioritising matches in part-of-speech and similarity of the headword, but the overall similarity of the entries is taken into account as well. For the older vocabularies the Latin equivalents are matched as well. The output is the input file, enriched with ID's of the chosen target(s), or with a code denoting that the entry should be established as a new lemma in the meta dictionary.

The “Konnektor” has been an invaluable tool in our linking tasks, but it is still too early in the process to evaluate both the input technique as well as the tool. This is something that will be examined in further detail when we have worked through a larger amount of material during the project.

Generation of keyword entries

The purpose of this task is to generate user-friendly, relatively short and plain entries that collect and condense the information found in a group of linked entries. The aim of these entries is not to present the dictionary content to the user, but simply to give the user an idea of the meaning of the word. Thus, we would ideally like to present a series of definitions or equivalents without evidence, sources, etymology, etc.

As mentioned above, the dictionary sources are quite heterogeneous, and that challenges the generation of the keyword entries. For this reason, we aim to generate content only where the result is meaningful. Thus, if a generated extract is empty, or if it is too complex, we suppress it, and the user will have to follow the link to the actual source entry.

Fig. 1 shows an example of a new entry (*belakke*, ‘defame, slander’, obsolete in modern Danish) where several definitions/equivalents are extracted. Fig. 2. shows an example (*afladsbrev*, ‘letter of indulgence’) where no meaningful content could be extracted, thus only presenting links to the source entries (in the yellow link box).

Fig. 1: An entry with generated content (and additional links to other dictionaries).

Fig. 2: An entry with no meaningful content for the time period, only links other dictionaries.

Concluding remarks

Currently the dictionary linking, as well as the content extraction and the construction of the website, is a work in progress. We anticipate that both the extraction process and the presentation of the dictionary will improve as soon as we receive feedback from the project's philologists and other users. Furthermore, it is our hope that the enrichment of the meta dictionary will enable us to exploit this data in future projects.

References

Nielsen, Marita Akhøj. *Hvorfor taler vi dansk? Om reformationen og sproget.* København: Eksistensen. 2017.

Svendsen, Mette Marie Møller, Nicolai Hartvig Sørensen & Thomas Troelsgård. "Superordbog og salmesang: ordbogslinkning i praksis" in: *Nordiske studier i leksikografi 15*. Helsinki: Nordisk forening for leksikografi. Expected time of publication: Late 2020.