WP 9 – infrastructure access at DSL

Resources, tools, infrastructure or research facilities to which DSL provide access

----

Tools

Tools for internal use at DSL, to be used for investigations on Danish texts and dictionaries:

- Corpus tool, statistic tool (Word2Vec model).
- Access to the xml editing system 'iLEX' and the Danish resources which are edited in this system at DSL.

Resources

A dictionary, a thesaurus and two computational lexicons for modern Danish sharing sense id numbers. The four resources are described in detail below. The first two are only available for research at DSL.

1) Traditional dictionary, modern Danish: The Danish Dictionary, DDO (ongoing project)

The DDO dictionary (*Den Danske Ordbog*) is a comprehensive monolongual dictionary of contemporary Danish, edited at a scholarly basis. It was originally published in print in 6 volumes in the years 2003-2005, but is nowadays published online at ordnet.dk/ddo (and also as app "Den Danske Ordbog"). Currently it involves seven editors/computational linguists and is being extended with the full description of 10,000 lemmas (2015-2018). The dictionary is edited in xml in a custom-designed structure and provides information on form, meaning and use of words belonging to the general vocabulary of Danish. The dictionary-making process is based on corpus inspection, and the development of the corpus as well as the corpus tools is part of the project. The dictionary covers 81,000 lemmas and 13,000 fixed expressions which are described by 120,000 sense definitions with identified genus proximum, and furthermore 14,000 lemmas without definition. The sense inventory, the fixed expressions, the collocations and the valency patterns have been used to compile a series of other lexical resources at DSL: the Danish Thesaurus and, together with CST/the University of Copenhagen, the two computational lexicons the Danish Wordnet *DanNet* and the Danish frame lexicon. All resources share sense id numbers with the dictionary. We aim at linking the DDO dictionary at either lemma and/or sense level to elder Danish dictionaries published by DSL, especially Dictionary of the Danish Language (ODS) and Old Danish Dictionary (GO). Based on the shared id numbers we recently identified and integrated relevant data on related words from the Danish Thesaurus into the dictionary entries. Both the linking from DDO to the other resources and its detailed xml structure which allows for the identification of very specific types of lexical information open up for many types of data combinations and lexical studies.

Published: online 2008 (ordnet.dk/ddo), in print 2003-2005
Language: modern Danish (1955-present) / general language
Type: monolingual, corpus-based, traditional dictionary
Data: xml-structure
Dictionary-making process: corpus, corpus tools, lemma selection, editing, digital publishing
Entries: 95.000 (of which 14,000 have no sense definition yet)
Fixed expressions: 13,000
Sense definitions: 120,000
Linked at sense level: Yes. The Danish Thesaurus, Danish WordNet, Danish Frame Lexicon
Visiting period: all year except 20 Dec. - 2. January, 1. July- 15. August
Currently involved editors: 7 lexicographers/computational linguists

Contact: Lars Trap-Jensen, ltj@dsl.dk, Henrik Lorentzen, hl@dsl.dk, Sanni Nimb, sn@dsl.dk, Thomas Troelsgård, tt@dsl.dk

2) Thesaurus, modern Danish: The Danish Thesaurus "Den Danske Begrebsordbog" (Ongoing project)

The Danish Thesaurus was published in print in 2015. It is based on the lemmas and fixed expressions in the DDO dictionary, but also on the many collocations described in the dictionary. It organises most of the word senses described in DDO in 22 chapters and 888 sections and presents the words and expressions in semantic groups with keywords. In each group the words are presented in semantic order. The underlying xml-document contains formal information on the semantic groups allowing for the identification of e.g. persons, acts etc. The thesaurus, which is edited in xml, is currently being extended with the DDO senses which are not yet integrated. We hope to be able to publish an online version of the book in the next years to come, depending on funding. The most relevant parts of the thesaurus sections have already been automatically identified and integrated in the DDO online dictionary in the form of related words for many senses, based on the shared id numbers in the two resources. Its organizing of the Danish vocabulary into annotated semantic groups was used to compile the Danish FrameNet lexicon, and the two resources share sense id numbers with DDO. The linking from the thesaurus data to the other resources opens up for many types of data combinations and lexical studies.

Published: in print 2015
Language: modern Danish (1955-present) / general language
Type: monolingual, corpus-based, traditional dictionary
Data: xml-structure
Dictionary-making process: lemma selection, editing, semantic annotation
Words and expressions: 204,000 (119,000 unique)
Linked at sense level: Yes. The Danish Dictionary (DDO), Danish WordNet *DanNet*, Danish FrameNet Lexicon
Visiting period: all year except 20 Dec. - 2. January, 1. July- 15. August
Currently involved editors: 2 lexicographers/computational linguists
Contact: Sanni Nimb, sn@dsl.dk, Thomas Troelsgård tt@dsl.dk

3) Computational lexicon, modern Danish: *The Danish FrameNet Lexicon* (also available for research outside DSL)

The Danish Framenet Lexicon was compiled 2016-2017 and describes 12,142 Danish lemmas with one or more frame values from the Berkeley FrameNet model. It furthermore gives information on the type of phrases and multiword units that would typically evoke the different frames of a lemma. It was compiled in 2016-2017 in collaboration with the University of Copenhagen on the basis of the vocabulary from the Danish Thesaurus, and aims at supplying semantic annotators of Danish texts with a reduced set of frame values, typically 3-4 per verb and 1-2 per verbal noun which are most likely to be relevant out of more than 1,000 values in Berkeley FrameNet when the text is to be annotated. 671 different frames were used to describe the lemmas which represents 80 % of the DDO dictionary. It is available as a comma-separated file at https://github.com/dsldk/.
The lexicon allows the study of different semantic groups of especially Danish verbs (and deverbal nouns), e.g. Danish motion verbs.

Published: 2017 https://github.com/dsldk/
Language: modern Danish (1955-present) / general language

Type: monolingual, corpus-based, computational lexicon
Data: comma-separated file (spreadsheet)
Dictionary-making process: linked data (thesaurus/dictionary), assignment of English frames
Entries: 12,142
Verbs: 5,300, Nouns: 6,490
Linked at sense level: Yes. The Danish Dictionary, The Danish Thesaurus, the Danish WordNet
Visiting period: all year except 20 Dec. - 2. January, 1. July- 15. August
Currently involved editors: 1 lexicographer/computational linguist
Contact: Sanni Nimb, sn@dsl.dk

4) Computational lexicon, modern Danish: The Danish WordNet *DanNet* (Ongoing project at the University of Copenhagen, also available for research outside DSL)

DanNet was compiled from 2004-2013 together with the University of Copenhagen. The compilation was based on the DDO lexical data and its sense inventory, including the sense definitions and especially the fact that the genus proximum of each sense is tagged in the xml structure of the dictionary.
The WordNet contains 65,000 synsets which are provided with an ontological type and a link to the closest hypernym. The synset members are linked to DDO senses. 5,000 Danish synsets are furthermore linked to the equivalent English synset in Princeton WordNet (by the relation eq_has_synonym), also labelled Princeton Core.

Published: 2013 http://wordnet.dk/lang.html
Language: modern Danish (1955-present) / general language
Type: monolingual, computational lexicon
Data:
Dictionary-making process:
Synsets: 65,000
Linked at synset member level: Yes. The Danish Dictionary, The Danish Thesaurus, the Danish WordNet
Visiting period: all year except 20 Dec. - 2. January, 1. July- 15. August
Currently involved editors:
Contact: Sanni Nimb, sn@dsl.dk, Nicolai H. Sørensen, nhs@dsl.dk

Other resources, also available for research outside DSL

- A number of Danish corpora and word lists http://korpus.dsl.dk/resources.html
- Manually POS-tagged corpus of Danish (PAROLE)
- SemDaX: manually tagged semantic corpus.

Research facilities / infrastructure

DSL provides access to two research infrastructures, one with the focus on dictionary editing and one with the focus on digitalisation of historic dictionaries

1: The process of editing and publishing corpus-based dictionaries on a scholarly basis; the linking of lexical resources

DSL is currently editing a modern Danish dictionary which is published online. The dictionary-making process includes all steps, from corpus creation, the creation of corpus tools and statistic investigations, the lemma selection based on corpus, the editing process based on corpus investigations, and finally the digital publishing. The sense inventory of the dictionary constitutes the skeleton to which all other DSL resources of modern Danish are linked by shared id numbers: the Danish thesaurus, the Danish WordNet and the Danish FrameNet lexicon. The thesaurus is currently being extended with senses from the dictionary in order to cover the full sense inventory. The linked data is already used for different purposes, one of which is the integration of related words from the thesaurus into the dictionary sense desriptions. In the ELEXIS project, we carry out research on how to link dictionaries of elder Danish to the modern resources.

2: Digitisation and online publishing of historic dictionaries

Since 2008 DSL has specialized in digitising and online publishing historic Danish dictionaries, as well as dictionaries that have been published by DSL only in print. In some cases the online publishing includes an app version. Two computational linguists are involved in this work. The first (and most comprehensive) online published Danish dictionary is still being developed with even more fine-grained tags based on the typography of the printed version. Also a new Swedish-Danish dictionary published in 2010 is to be published online in 2018 (and we plan to publish the Danish Thesaurus in the years to come, probably before 2021, depending on funding). The dictionaries are listed here:

Dictionary of the Danish Language (ODS)	1918-1956	2005, also app	Danish (1700-1955)
Meyer's Loanword Dictionary	1837/.../1924	2014, also app	Loanwords in Danish
Moth's Dictionary	169x-1718	2013, also app	Danish around 1700
Holberg Dictionary	1981-1988	201?, also app	Danish (in the works of L. Holberg, 1684-1754)
Dictionary of older Danish (Kalkar's Dictionary)	1885-1918	2017	Middle Danish and early modern Danish (1300-1700)
Swedish-Danish Dictionary	2010	To be published in 2018	Swedish-Danish
Jensen & Goldschmidt	1886/1920	2017	Latin-Danish

Data: xml-structure
Dictionary-making process: digitisation, digital publishing
Visiting period: all year except 20 Dec. - 2. January and 1. July - 15. August
Currently involved editors: 2 computational linguists
Contact: Thomas Troelsgård tt@dsl.dk

Three of the dictionaries are described in more detail below.

1) Dictionary of the Danish Language, ODS (Ongoing project: adding tags to xml)

Ordbog over det danske Sprog (Dictionary of the Danish Language) is an online dictionary (also as app) based on a printed dictionary which was published in the years 1918-1956. The 28 volumes were completed within 40 years and later increased by 5 supplementary volumes (published 1992-2005). Recently also the supplementary volumes have been integrated in the online version. It is the largest monolingual dictionary compiled for Danish. The dictionary provides information on form, meaning and use of words belonging to the general vocabulary of Danish, and is meant to be useful for a wide range of users. It is rich on examples, primarily from literature. The data originate from a digitisation based entirely on typography. Currently it involves 2 computational linguists who are adding tags describing the lexicographic structure (preserving the original markup for historical reasons). As a result, in the current data some information types are well recognised and tagged, while others are not.

Published: online 2005 (ordnet.dk/ods), in print 1918-1956 (suppl. 1992-2005)
Language: Danish (1700-1955) / general language
Type: monolingual, traditional dictionary
Data: xml-structure (digitised elder manuscript)
Dictionary-making process: digitalisation from typography, adding tags, digital publishing
Entries: 210,000
Linked at sense level: No. The future plan is to link its lemmas and senses to DDO and historical dictionaries of Danish
Visiting period: all year except 20 Dec. - 2. January, 1. July- 15. August
Currently involved editors: 2 lexicographers/computational linguists
Contact: Sanni Nimb, sn@dsl.dk, Thomas Troelsgård tt@dsl.dk

2) *Moths Ordbog* (Moth's Dictionary)

Moths Ordbog (Moth's Dictionary) is an online dictionary (also app) based on a digitised version of a handwritten manuscript from the late 17th/early 18th century. Most entries have a Latin equivalent for the Danish headword, and about half of the entries also have Danish definitions. There is information about inflection, and often usage examples, including multi-word expressions (also with Latin equivalents). The current structure is an original typographic markup supplied with tags describing the lexicographic information types.

Published: online 2013 (http://mothsordbog.dk), handwritten document 169?-1718
Language: Danish (around 1700), Latin / general language (and some specific language (medicine / plants))
Type: historical, bilingual, traditional dictionary
Entries: 105,000 (~81,000 full)
Data: xml-structure (digitised handwritten manuscript)
Dictionary-making process: digitalisation from handwrite, adding tags, digital publishing
Linked at sense level: No. Future plan is to link to ODS and historical dictionaries of Danish
Visiting period: all year except 20 Dec. - 2. January, 1. July- 15. August
Currently involved editors: 2 computational linguists
Contact: Thomas Troelsgård tt@dsl.dk*Kalkars Ordbog* (Kalkar's Dictionary - Dictionary of older Danish)

3) *Kalkars Ordbog* (Kalkar's Dictionary)

Kalkars Ordbog (Kalkar's Dictionary) is a digitised version of a five volume dictionary of Middle Danish and early modern Danish, published in the period 1881-1918. The structure is nested with derivatives and compounds placed under a simplex headword, including multi-word expressions (also with Latin equivalents). The data originate from a digitisation based entirely on typography. The current data structure has been supplied with tags embedding some of the lexicographic information types.

Published: online 2017 (http://kalkarsordbog.dk), in print 1881-1918
Language: Middle Danish and early modern Danish (1300-1700), Latin/ general language
Type: historical, monolingual/bilingual, traditional dictionary
Data: xml-structure (digitisation based entirely on typography)
Dictionary-making process: digitalisation, adding tags, digital publishing
Entries: 30,700 (24,700 full)
Linked at sense level: No. The future plan is to link to ODS
Visiting period: all year except 20 Dec. - 2. January, 1. July- 15. August
Currently involved editors: 2 lexicographers/computational linguists
Contact: Sanni Nimb, sn@dsl.dk, Thomas Troelsgård tt@dsl.dk