Combining Dictionaries, Wordnets and other Lexical Resources– Advantages and Challenges

Bolette Sandford Pedersen[1], Sanni Nimb[2], Sussi Olsen[1], Nicolai Hartvig Sørensen[2]
University of Copenhagen[1], The Danish Society of Language and Literature[2]
Njalsgade 139, DK-2300 S[1], Christians Brygge 1, DK-1219
bspedersen@hum.ku.dk, sn@dsl.dk, saolsen@hum.ku.dk, nhs@dsl.dk

Abstract

In this paper we account for the advantages, challenges and pitfalls that we have encountered when compiling language technology (LT) resources based on dictionary information and vice versa. We describe the main lines in our collaborative work during the last decade and based on this experience, we provide some suggestions and recommendations in order for dictionaries to become more standardised and multifunctional and thereby also more directly useful for LT.

Keywords: lexical resources, wordnets, framenets, annotated corpora, language technology, international standards, language transfer

1. Compiling LT resources from dictionaries and vice versa

In this paper we account for the advantages, challenges and pitfalls that we have encountered when compiling language technology (LT) resources based on dictionary information and vice versa. Our focus is on a medium-resourced language, namely Danish, where LT resource scarcity has prompted us to look seriously into the perspective of re-using existing lexical resources.

To this end, it is important to stress that dictionaries are not just systematic collections of words with information about morphology and syntax; they are cultural testimonies in the sense that they describe the society and culture in which they are being compiled. Ideally, the LT systems that we develop for use in both our private and professional lives should reflect the same dimensions. However, if we solely adapt our future LT systems on the basis of English language models, there is a danger that this dimension is completely overlooked.

In order to address this challenge, the Danish language and language technology community has in recent years focused on methods for building language technology resources that:

- employ existing high-quality lexical data of Danish,
- comply with international standards, *and*
- incorporate elements of language transfer from better resourced languages where relevant[1]

In addition to this combination of approaches, focus has been into keeping a reference point across all the developed resources in terms of common sense identifiers or a common "core" so to speak. This approach has enabled the teams to not only produce LT resources from traditional dictionary work, but also go the other way: To exploit LT resources when developing a new Danish thesaurus.

Where a close collaboration between a dictionary publisher and a university institute (as seen in our case between The Society for Danish Language and Literature and the Centre for Language technology at the University of Copenhagen), is not seen so often, the idea of developing lexical cores as a basis for new resources, is not a new or unique approach. Examples are such as The DANTE database (Atkins 2010) which is a lexical database which provides a fine-grained, corpus-based description of the core vocabulary of English. SALDO (Borin et al. 2013) is a Swedish semantic and morphological lexical resource primarily intended for use in LT applications, which however, is closely entangled with two paper dictionaries as well as with the Swedish wordnet. Similar to SALDO, Cornetto stands for Combinatorial and Relational Network as Toolkit for Dutch Language Technology and is a lexical semantic database that combines a wordnet with framenet-like information for Dutch (cf. Vossen et al. 2013). The combination of the two lexical resources (the Dutch wordnet and the Referentie Bestand Nederlands) is claimed to provide a richer relational database to be used in LT.

Our own starting point for the collaborative work between resources, which has been realised for more than a decade, is the monolingual dictionary *Den Danske Ordbog* (DDO) and the Danish wordnet, DanNet; the latter compiled a decade ago with DDO as its primary source (Pedersen et al. 2009), but still complying with wordnet standards (Fellbaum 1998, Vossen 1999). To compile the wordnet we used a bottom-up strategy based on the hypernym given for each sense definition in the dictionary expressed in a specific genus proximum field. As consequence of this compilation approach, the two resources are linked at

[1] See for instance Pedersen et. al. (2018) for transfer of frame-semantic information from English.

sense level, allowing for the combination of all types of information across the two resources.

For instance, the links have been used to enrich the online version of the DDO, enabling users to browse related words in terms of hyponymy (Sørensen & Trap-Jensen 2010). The exact order of the hyponyms in the online presentation 'Beslægtede Ord' (Related Words, available 2009-17) was based on a calculation of semantic relatedness depending on information in the wordnet: a set of semantic relations and the ontological types. Another direct use of the combined data is the graphical representation of DanNet's hierarchies and relations at *andreord.dk* where the (restricted) definitions of DDO as well as domain information and citations from the dictionary are included. In Section 2 we describe the common sense inventory in more detail.

Most recently, the linked data has furthermore resulted in new resources in terms of an *annotated corpus*, *a Danish thesaurus* and a *Berkeley-style frame-lexicon* all of which we briefly account for in Section 3.

In Section 4 we sketch out some recommendations for a future larger degree of multi-functionality in the next generation of dictionary projects. In particular, we discuss the perspectives of future, truly digitally born lexical resources which are not limited or influenced by (former) physical issues, and which can therefore be compiled and interlinked with a higher degree of consistency.

2. One sense inventory as a common reference point

The DDO is corpus-based and continuously being extended with new words and senses. Entries are organized in main and sub-senses in a structure which to a high degree reflects the logical relations between a core sense and its either narrower or broader sense derivations as well as metaphorically derived senses. However, this general principle is sometimes downgraded for communicative purposes. For instance, very deep sub-sense structures are avoided, and very frequent senses have instead been upgraded to main senses, no matter whether there exists a logical relation to a core sense or not. What is also important to notice is that the first edition of DDO was published in print in six volumes. This influenced to a very high degree the sense structure of less frequent words. For such words the core and sub-senses were often merged into one definition in order to save space for a more detailed description of the very frequent words. Furthermore, many cases of regular polysemy are implicit in the dictionary, covered by only one sense.

When we compiled the Danish wordnet, DanNet (Pedersen et. 2009) from the DDO in a semi-automatic fashion, these informal deviations from the general structure caused some extra adjustment work in terms of reorganization of senses and collapses of some senses into the same synsets. Likewise, the adjustment and reorganization of the implicit DDO hyponymy structure was somewhat time consuming. For instance, we realized that many of the hyponymies found in the DDO had incorporated a great mixture of *natural* and *functional* kinds in Cruse's terminology (Cruse 2000), mixing natural taxonomies with layman's view of the concept's *function*. For instance, edible plants could have either 'plant' or 'vegetable' as their hypernym in the DDO depending somewhat on the lemma's frequency in the corpus and on its subsequent allotted physical space and unfolding in the original dictionary.

3. Developing new resources based on DDO/DanNet

3.1 Combinations of information from wordnet and dictionary: A thesaurus and a Frame lexicon

The semantic links between DanNet and the DDO further facilitated the compilation of a comprehensive thesaurus for Danish (Nimb et al. 2014 a; Nimb et al. 2014 b). Large hierarchies of words (i.e. all furniture or clothes), including links to the corresponding DDO senses, were directly transferred to the relevant thesaurus chapters. Data extracted from DDO in the form of definitions and synonyms was used to arrange the hyponyms into subgroups, and the categorization of senses profited from our experiences with the wordnet compilation.

Several of the semantic relations from DanNet were adapted in order to structure the thesaurus XML manuscript. By use of these formal semantic criteria, the vocabulary was annotated with core semantic types such as acts, events, properties, persons, artifacts etc., enabling us to keep track of the semantic grouping of words throughout the thesaurus project as well as to identify and extract precisely restricted semantic groups from the finished manuscript. In this way, approx. 1/5 of the words and expressions in the thesaurus were identified as acts or events and subsequently used for starting up the Danish frame lexicon. See Nimb et al. (2017) and Nimb (2018) for more details.

The chapter division in the thesaurus made it possible to identify precise semantic domains such as acts of 'communication' and 'cognition' and thereby to assign the appropriate frame in Berkeley FrameNet covering these exact domains to a large quantity of lexical units at a time. The resulting frames have been tested on restricted corpus data (Nimb et al. 2017), and the project has afterwards been extended in order to compile frames for the entire Danish act/event vocabulary. In a future project, we plan to study whether the sense links between the frame data and DanNet can be used to extend the wordnet with framenet information, i.e. especially to improve the verb hierarchies of DanNet.

3.2 A semantically annotated corpus

The common backbone sense inventory was also further exploited for annotating a corpus – annotations which were subsequently used for training a Danish sense tagger (Martinez et al. 2015 and Pedersen et al. 2018). Hence, the so-called SemDaX corpus (Pedersen et al. 2016) contains about 100,000 words with semantic annotations of varying granularity, annotated by humans. The most coarse-grained sense annotations are annotations of all content words with so-called *supersenses*, derived from Princeton WordNet's lexicographical files.

In addition to the supersense annotations, SemDaX comprises lexical sample annotations for a small set of highly ambiguous nouns. The fine-grained annotations are based on the set of senses in DDO. Each noun has been annotated with the full DDO sense inventory as well as with two different automatically clustered sense inventories of different granularity (Pedersen et al. 2018) based on their ontological type in DanNet.

All manual annotations were carried out in the annotation tool WebAnno (Yimam et al., 2013). The aim of the corpus is to serve as training and test data for word sense disambiguation, as well as to estimate the usefulness of the different sense annotation schemes by analyzing the data and the inter-annotator agreement.

4. Future dictionaries: How can they become more suitable for multiple purposes?

The Danish lexical core approach was initiated with the combination of a dictionary and a wordnet based on the common sense inventory. This initiative gave interesting insights and results and led on to other lexical products as described in the above. To sum up, the linked data combining hierarchical information, semantic relations, dictionary definitions, and dictionary synonyms has enabled us to compile a thesaurus and consequently also a frame lexicon in a very efficient way. The logical information from the dictionary sense structure combined with the ontological information in the wordnet has furthermore allowed us to carry out several comparative annotation studies with both full sense inventories and sense *clusters*. Using this corpus for word sense disambiguation has given us insights wrt. how to identify the most adequate levels of sense granularity – both for human annotators and for automatic systems.

The work has further provided insights into where dictionaries for human users lack explicit information which is needed for human language technology. One example is the logical relation between senses which should preferably be more specific and for instance described by more specific links. Another is the discrepancies in hypernym structure where space issues in the printed dictionary to some extent influenced the structure so that for instance regular polysemous lemmas did not systematically refer to their correct hypernyms.

Also the assignment of very coarse-grained semantic information, such as whether the sense is a first, a second or a third order type of entity (cf. Lyons 1977) would be very useful to have implicitly expressed in dictionaries, preferably by the use of simple attributes. Often dictionary definitions use polysemous words across the three semantic classes (i.e. figurative, abstract words that also have a concrete sense). This has as consequence that it is not at all easy to extract whether a standalone definition defines something concrete or abstract – or maybe even covers both cases – without having to look deeper into citations, other senses of the word etc. The same goes for many cases of regular polysemy. Precise attributes on regular polysemy patterns should preferably be included in dictionaries, allowing the editor to check out and mark which of the regular senses are accounted for in the description, based on corpus inspection.

Our work with dictionaries in an LT context has also inspired us the other way around regarding which supplementary information types seem useful for LT resources and have not previously been fully acknowledged as such. Surprisingly enough, for instance, the *function* relation (labelled the 'telic role' in Pustejovsky 1995, and 'functional/nominal' kinds by Cruse 2000) receives very little attention in the wordnet literature, and only very few wordnets contain – to our knowledge – this information type even if it proves quite crucial in many inference tasks in particular when it comes to tasks involving artifacts. The relation is highly represented in many DDO definitions where a concept's function is very often described – and when it is not, the integration with other resources is much more complicated. In fact, in Nimb & Pedersen 2000 we concluded that a concept's function often constitutes the very core of the figurative sense of the same word[2]. To this end, we would recommend that also this relation becomes formally explicit via the logical relations between senses as well as the function role formally explicit in dictionaries.

With regards to sense structure, one can only hope that future digitally born dictionary versions (where physical limitations is no longer an issue), will by and by result in a more consistent sense description where lesser frequent words are treated with same consistency as frequent words. Combined with a higher level of standardization – in our case partly introduced via the international wordnet and framenet standards – some of the obstacles that we have encountered in our work can hopefully gradually be reduced. However, there is no doubt that it requires explicit focus.

In fact, the newly embarked ELEXIS infrastructure has

[2] For instance, the telic role of *window*, namely to give access to a broader view of the surroundings from the inside of something, determines the figurative sense in a phrase like *a window to the world*.

exactly the goal of explicitly addressing cooperation and information exchange among lexicographical and LT research communities. The aim is to achieve a higher degree of standardisation and inter-functionality of existing and future dictionaries. The infrastructure is a newly granted project under the Horizon 2020 INFRAIA call, and the plan is to work with strategies, tools and standards for extracting, structuring and linking of lexicographic resources.

5. Bibliographical References

Atkins, B. T. S. (2010). The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data. In : G.-M. de Schryver (ed.) *A Way with Words: Recent Advances in Lexical Theory and Analysis*. A Festschrift for Patrick Hanks. Kampala: Menha Publishers.

Borin, Lars, Markus Forsberg, Lennart Lönngren (2013). SALDO: a touch of yin to WordNet's yang. In : *Language Resources and Evaluation, Volume 47, Issue 4*, pp 1191–1211.

Cruse, D.A (2000). *Meaning in Language*. Oxford: Oxford University Press.

DDO = *Den Danske Ordbog*. (E. Hjorth et al). 2003-2005. Det Danske Sprog- og Litteraturselskab & Gyldendal, Copenhagen.

Fellbaum, Christiane (ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT press.

Lyons, John (1977). *Semantics*. Cambridge University Press

Martínez Alonso, Héctor; Anders Johannsen; Sussi Olsen; Sanni Nimb; Nicolai Hartvig Sørensen; Anna Braasch; Anders Søgaard; Bolette Sandford Pedersen. (2015). Supersense tagging for Danish. In : *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015*, Linköping Electronic Conference Proceedings #109, ACL Anthology, Linköping University Electronic Press, Sweden.

Martínez Alonso, Héctor; Anders Johannsen; Sanni Nimb; Sussi Olsen; Bolette Sandford Pedersen. (2016). An empirically grounded expansion of the supersense inventory. In : *Proceedings of Global Wordnet Conference 2016*.

Nimb, Sanni (2018). The Danish FrameNet Lexicon: method and lexical coverage. In : *Proceedings from the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons.* Miyazaki, Japan

Nimb, Sanni; Braasch, Anna; Olsen, Sussi; Pedersen, Bolette Sandford; Søgaard, Anders. (2017). From Thesaurus to FrameNet. In: *Proceedings of eLex 2017*, Leiden.

Nimb, S. & B.S. Pedersen (2000). Treating Metaphorical Senses in a Danish Computational Lexicon - Different Cases of Regular Polysemy. In : *Proceedings from The Ninth Euralex International Congress pp. 679-691*, Universität Stuttgart Germany.

Nimb, Sanni, Henrik Lorentzen, Liisa Theilgaard, Thomas Troelsgård, Lars Trap-Jensen (2014 a). *Den Danske Begrebsordbog*. Det Danske Sprog- og Litteraturselskab og Syddansk Universitetsforlag

Nimb, Sanni, Lars Trap-Jensen, Henrik Lorentzen (2014 b) The Danish Thesaurus: Problems and Perspectives. In: Andrea Abel, Chiara Vettori & Natascia Ralli (eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: EURAC Research, pp. 191-199

Pedersen, Bolette S., Manex Agirrezabal, Sanni Nimb, Sussi Olsen, Ida Rørmann (2018). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In: *Proceedings of GWC2018*, Singapore.

Pedersen, Bolette S., Sanni Nimb, Jørg Asmussen, Nicolai Sørensen, Lars Trap-Jensen, Henrik Lorentzen. (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. In: *Language Resources and Evaluation*, Computational Linguistics Series, pp.269-299.

Pedersen, Bolette S., Sanni Nimb, Anders Søgaard, Mareike Hartmann, Sussi Olsen (2018). A Danish FrameNet Lexicon and an Annotated Corpus Used for Training and Evaluating a Semantic Frame Classifier. In: *Proceedings of LREC 2018*, Miyazaki, Japan.

Pedersen, Bolette S.; Braasch, Anna; Johannsen, Anders Trærup; Martínez Alonso, Héctor; Nimb, Sanni; Olsen, Sussi; Søgaard, Anders; Sørensen, Nicolai. (2016). The SemDaX Corpus - sense annotations with scalable sense inventories. In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*. Portorož, Slovenia.

Pustejovsky, James (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA.

Sørensen, Nicolai Hartvig. & Trap-Jensen, Lars (2010). Den Danske Ordbog som begrebsordbog. In : Harry Lönnroth, Kristina Nikula (eds.) *Nordiska Studier i Leksikografi 10*, NFL-skrift nr 11, Tammerfors 2010, pp. 164-179.

Vossen, P. (ed). (1999). *EuroWordNet : A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, The Netherlands.

Vossen, P., I. Maks, R. Segers, H. van der Vliet, M. Moens, K. Hofmann, E. Tjong Kim Sang, and M. de Rijke (2013). Cornetto: a lexical semantic database for Dutch. In : *Essential speech and language technology for Dutch, results by the Stevin-programme*, P. Spyns and J. Odijk, Eds., Springer series Theory and Applications of Natural Language Processing, 2013, pp. 165-184.

Yimam, S.M., Gurevych, I., Eckart de Castilho, R., & Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In: *Proceedings of ACL-2013*, demo session, Sofia, Bulgaria.