Merging a Syntactic Resource with a WordNet - a Feasibility Study of a Merge between STO and DanNet

Bolette S. Pedersen, Anna Braasch, Lina Henriksen, Sussi Olsen, Claus Povlsen
Centre for Language Technology, University of Copenhagen
Njalsgade 80, DK-2300 S
E-mail: bolette@cst.dk, anna@cst.dk, lina@cst.dk, sussi@cst.dk, claus@cst.dk

Abstract

This paper presents a feasibility study of a merge between *SprogTeknologisk Ordbase (STO)*, which contains morphological and syntactic information, and *DanNet*, which is a Danish WordNet containing semantic information in terms of synonym sets and semantic relations. The aim of the merge is to develop a richer, composite resource which we believe will have a broader usage perspective than the two seen in isolation. In STO, the organizing principle is based on the observable syntactic features of a lemma's near context (labeled syntactic units or SynUs). In contrast, the basic unit in DanNet is constituted by semantic senses or - in wordnet terminology - synonym sets (synsets). The merge of the two resources is thus basically to be understood as a linking between SynUs and synsets. In the paper we discuss which parts of the merge can be performed semi-automatically and which parts require manual linguistic matching procedures. We estimate that this manual work will amount to approx. 39% of the lexicon material.

1. Introduction: the need for a composite lexical resource

During the last decade, two large computational lexicon resources have been developed for Danish: *SprogTeknologisk Ordbase (STO)*, which contains morphological and syntactic information, and *DanNet*, which is a Danish WordNet containing semantic information in terms of synonym sets and semantic relations. Each of these computational resources fills an important gap in the development of Danish language technology. However, although we are only just now completing the first phase of 40,000 synsets in DanNet, a comparison of the two resources has recently been initiated with the aim of developing a composite resource covering morphology, syntax and semantics.

A combination of the semantic descriptions in DanNet with the morphological and syntactic descriptions of STO will result in a much richer resource than the two seen in isolation, and it will undoubtedly constitute a very strong lexical resource for language technology applications in future, such as systems for syntactic and semantic text mark-up, disambiguation systems, text generation systems, machine translation systems as well as systems for advanced information retrieval.

This paper presents the results of a feasibility study of such a merge.

2. Related work

Combining computational lexicons with different information types is not a new idea. Actually, in the STO project, which is based on the PAROLE/SIMPLE lexicon model (Ruimy et al. 1998, Lenci et al. 2001), a mapping between a syntactic and a semantic level of representation was foreseen already in the data model. In the PAROLE/SIMPLE projects, a small Danish semantic lexicon was developed and partly interlinked with the syntactic part (Pedersen & Paggio 2004). The semantic apparatus of PAROLE/SIMPLE builds on Pustejovsky's Generative Lexicon (Pustejovksy 1995) and is very rich and complex in its composition.

However, exactly this feature of complexity made it more or less unrealistic for us in the given situation to scale-up the semantic coverage of the resource into something practically useful. In contrast, DanNet contains a much leaner semantics (following the EuroWordNet Framework, cf. Vossen 1999) which is furthermore semi-automatically processed on the basis of an existing lexicon, Den Danske Ordbog (DDO = Hjorth et al. 2005), which contains explicit and extractable genus proximum and synonymy information (cf. Asmussen et al. 2007, Pedersen & Nimb 2008).

Also in the wordnet community the interest of enriching wordnets with morphological and syntactic information is increasing. If wordnets are to be used in general language technology environments as mentioned above, the need for syntax and morphology proves obvious. The newly initiated Cornetto project (Vossen et al. 2007) is an attempt of merging Dutch WordNet with a morphosyntactic resource of Dutch realised in terms of a FrameNet-like structure. Similar to the aim of the present project, the goal of Cornetto is to attain a resource that combines semantic, formal semantic and combinatoric information in order to achieve better resources for natural language engineering technologies. In contrast to the STO-DanNet merge where STO does not establish word meanings but rather syntactic patterns, the Cornetto project deals with the task of merging two resources with different approaches to word meaning.

Closer to tasks of the STO-DanNet merge are the initiatives regarding a merge of ItalWordNet and PAROLE-SIMPLE-CLIPS (Roventini & Ruimy 2006). The two resources to be merged in this initiative are actually built on the same two bases as STO and DanNet, respectively, namely the PAROLE standards (Ruimy et al. 1998) and EuroWordNet (Vossen 1991). One main difference, however, is that the Italian SIMPLE lexicon is much further developed than the Danish one, and therefore the merge in this project - like in the Dutch one - takes place between two interpretations of semantic meaning.

A fact to be noted regarding these projects is that they are described as *new* resource building projects. In other words, the practical work of merging resources built on different linguistic grounds is considered a substantial task that can be compared with building a new resource.

3. Examining the compatibility between STO and DanNet

3.1 What to be merged

In STO the organizing principle is based on the observable syntactic features of a lemma's near context(s) (cf. Braasch & Olsen 2004). The definition of a syntactic unit is based on a combination of the complementation properties and a number of other features (e.g. reflexivity, control and raising) described by attribute/value-pairs. The combination of the lemma and one of its syntactic patterns make up the syntactic unit (SynU), and a lemma may have one or more such units, depending on the syntactic constructions the lemma appears in. This means that homonyms that share syntactic behaviour are represented in one and the same SynU regardless of their differing senses.

In contrast, the basic unit in DanNet is constituted by the synonym sets (synsets). A synset is defined as the set of lexical units that refer to the same concept; the most prototypical case being, however, that a synset constitutes *one semantic sense* of a given lemma. The merging of the two resources is thus basically to be understood as a linking between SynUs and synsets.

3.2 Size of the merge task

The merging of syntactic and semantic information for nouns is in this context to a great extent workable by means of (semi-)automatic processes because of the less complex nature of their syntactic properties. To be more precise, STO contains 33,000 nouns, of which 23,500 have only a single avalent reading. This leaves us with a set of less than 13,000 valent noun readings which require closer human inspection, as described above.

In the case of verbs, STO includes 5,500 verbs amounting to 8,500 different syntactic readings, i.e. less than two SynUs per verb to be considered. At the semantic level the structure generally proves to be much more complex. As will be exemplified in the following, we foresee that almost all verbs to be merged will require manual procedures.

Regarding adjectives, DanNet currently contains only a small set of approx. 1,000 synsets. Since the sense establishing structure of this word class is not yet fully settled upon, the present feasibility study contains no considerations regarding the merge of this word category.

3.3 Apparatus for defining merge types

The result of the merge will be a new resource with STO-information forming the basis, and DanNet-information constituting a semantic enrichment. This merge will be based on SynU/synset merges according to two overall methods: the *simple* and the *composite merge* types. The simple merge refers to one-to-one relations, whereas the composite merge involves more than one unit from one or both resources. This type of merge therefore refers to one-to-many, many-to-one and many-to-many relations.

Relations established between units may be of different kinds reflecting varying degrees of equivalence:

- *Simple/Composite Equivalence* – the elements of the merge are fully consistent, i.e. one/a set of STO-SynUs is consistent with one/a set of DanNet-synsets
- *Simple/Composite Similarity* – one/a set of STO-SynUs is not completely consistent with one/a set of DanNet-synsets, but the concepts reflected by the SynUs/synsets share strong similarities
- *Simple/Composite Subset* – one/a set of STO-SynUs covers a subset of one/a set of DanNet-synsets (it has been decided not to include a relation reflecting DanNet-synsets covering only a subset of STO-SynUs as these merged entries will not be incorrect, merely incomplete).

Lemma	**SynUs in STO**	**Synsets in DanNet**	**Relation type**	**Merge**
mål (noun)	**1**: monovalent, obligatory N (type: *specifier)* **2**: monovalent, optional genitive N **3**: monovalent, optional PP (*for*)	**A**: destination **B**: aim, objective **C**: goal *(sports)* **D**: dimension(s) **E**: target *(military)* **F**: target **G**: scoring **H**: measure **I**: size **J**: quantity *(measure)* **K**: portion *(measure)* **L**: language	Composite similarity	1 – J+K 2– A+B +(C)+D +E+(F) +G+(H) +I+L 3– A+B+E

Table 1: Merging STO's SynUs of *mål* (destination, goal..) with its synsets in DanNet

3.4 Merging data

The experimental data set has been examined with two aims. First, we have wanted to detect the degree of possible coincidences in the distribution of syntactic behaviour as opposed to semantic descriptions in terms of synsets or senses. Secondly, we have wanted to achieve a general overview of the feasibility and complexity of the merging task by identifying typical problem areas as well as appropriate solutions to these.

The experimental data set is composed of 100 nouns and 20 verbs of various types, both monosemous and polysemantic lemmas having one or more syntactic complementation patterns.

Table 1 shows an example of a typical merging situation illustrating one of the challenges met in the project. The noun *mål* has 12 synsets (see column 3) and all of them are covered by the construction types listed in STO (column 2).

In addition, C, F, and H expose a typical merge problem regarding nouns: The parentheses in column 5 indicate that these senses can *only* occur as zerovalent, and this structure is not explicitly covered by the STO descriptions. Therefore, the relation cannot be regarded as a case of purely *Composite Equivalence*. On the other hand, the zerovalent structure is covered by the monovalent structure with an optional, genitive noun complement. We therefore label the relation of the type *Composite Similarity*.

The case is typical of noun complementation patterns because of the encoding strategy employed in STO which is based on the general principles of *optionality* and *broad syntactic description*s. This means that practically all complements of nouns are considered optional. As a consequence a monovalent description with an optional NP or PP complement also covers constructions without a complement regardless of whether the construction is zerovalent *or* an elliptic construction with the complement omitted.

In order to increase the merge precision, we conclude that the broad syntactic descriptions need to be unfolded in two separate descriptions. Such an unfolding of compact descriptions can be done automatically and will radically increase the number of the more manageable relation type: *Composite Equivalence* (as would be the case for *mål*).

As already mentioned, STO contains 8,500 syntactic verb readings derived from 5,500 verbs. An example of a verb merge is given by the verb *forsøge* in Table 2. For verbs it is worth noting that all complements are considered obligatory unless they are explicitly marked as optional – opposite to the encoding strategy applied for noun complementation.

The lemma has four syntactic constructions in STO and four main senses in DanNet. This does not necessarily indicate a *Simple Equivalence* or a 1:1 relationship between the syntactic descriptions and synsets. One of the senses (A) relates to two different syntactic constructions (1+2), whereas two of the senses ((B + C) map onto one and the same syntactic description (4). Only sense (D) and the syntactic description (3) show the 1:1 relationship. Sense B: ‘to put forward (*cautiously)*’ has a characteristic feature which classifies this mapping as ‘not completely consistent’, since the oblique object (*med ngt* – with something) in this sense is facultative, but however always semantically implied. In contrast, there is a strong similarity between the syntactic construction of sense C since the oblique object of this sense is obligatory. Such cases are therefore regarded as *Composite Similarity*.

As already seen, a frequent merge task includes cases where a SynU covers a set of different senses. In some cases, these can be terminologically differentiated as in examples (1) and (2) below which relate to the domains of law and sports, respectively:

(1) De <u>dømte</u> Dr. La Coste skyldig
(They <u>convicted</u> Dr. La Coste guilty)

(2) Dommeren <u>dømte</u> bolden ude
(The referee <u>called</u> the ball out)

This terminological difference evokes two senses in DanNet as seen in Table 3, sense A and B, although they are syntactically similar. Likewise, a *figurative* sense calls for yet another synset which may however very well be expressed in the same way syntactically, as seen in example 3.

Lemma	SynUs in STO	Synsets in DanNet	Relation type	Merge
forsøge (verb)	**1**: divalent, obl. PP (prep=*på* (on, at) + infinitive clause with subject control) **2**:divalent, obl. NP or interrogative sentence **3**: divalent, + reflexive; introducer=*som* (as) +NP **4**: divalent, +reflexive; PP (prep=*med* (with) + NP or infinitive clause with subject control)	**A**: to attempt/try **B**: to put forward (*cautiously)* **C**:to endeavour /try *(intensively)* **D**: try one’s hand at	Composite similarity	1+2 – A 3 – D 4 – (B)+C

Table 2: Merging STO’s SynUs of *forsøge* (attempt, endeavour ..) with its synsets in DanNet

Lemma	**SynUs in STO**	**Synsets in DanNet**	**Relation type**	**Merge**
dømme (verb)	**1**: trivalent, obligatory object + attribute to the object **2:** trivalent, obligatory object + PP (to) PP-complement: infinitive with object control **3:** trivalent, obligatory object + PP (for) PP-complement: infinitive with object control	**A:** pronounce a sentence in court (*law*) **B:** to decide in a sports match (*sports*) **C*:*** decide for sby to receive a specific (rough) treatment (*figurative*) **D:** evaluate, estimate (*figurative*)	Composite similarity	1–A +(B + D) 2+3 – A+C

Table 3: Merging STO's SynUs of *dømme* (judge, evaluate, call ..) with its synsets in DanNet

(3) De dømte slaget til at være tabt
lit: They estimated the battle lost

Apart from these very general observations regarding discrepancies between syntactic and semantic descriptions, several other potential problem cases in relation to verbs have been detected during the feasibility study, such as:

- *Phrasal verbs and other verb constructions*: These tend to cause problems because of their productivity: In other words, we found several cases of disagreement between the two resources regarding coverage (one resource has one set of particles encoded, the other has another).

- *Coverage*: It is not always obvious *which* syntactic patterns are actually covered by a given synset in DanNet since a synset is often underspecified in this respect.

- *Synsets with multiples lemmas*: In cases where a synset encompasses more than one lexical unit, the merge problems escalate; for instance in the case of 'prepare food' which can be lexicalised in Danish by several lemmas: *tilberede, tillave* and *lave (mad)* each of which has its own set of syntactic patterns to be compared with.

4. Encoding tool

On the basis of the figures given in Section 3, we estimate that approx. 61 % of the vocabulary can be automatically merged. The rest of the vocabulary will need to be merged manually. To ease the manual work and increase efficiency we plan to develop an encoding tool (most presumably in the programming language Python in which the DanNet encoding tool is written).

The tool is intended to include the following basic functionalities:

- For a given lemma, a presentation of the list of SynUs for the user to take into account.

- For the given set of SynUs to a lemma, a presentation (in a split screen-like fashion) of the relevant list of synsets in DanNet including gloss and ontological type.

- An easy-click facility which enables the user to establish links between matching SynUs and synsets in a flexible way.

- A facility which automatically stores the enriched STO resource in the relevant database format (Oracle).

The tool will include a wizard-like function where - once the word class or sub-wordclass has been selected - the user is automatically presented to the next word on the STO wordlist to be enriched with semantic information.

In addition, the tool will facilitate the merge of more than one lemma from STO with the same synset in DanNet in cases of synsets with more than one lemma (synonymy).

5. Conclusion

One conclusion that can be drawn on the present feasibility study is that the merge of syntactic and semantic information for nouns is to a great extent workable by means of (semi-)automatic processes because of the less complex nature of their syntactic properties. In the case of verbs, however, the structure generally proves to be much more complex as been exemplified in Section 3. On the basis of the problem cases described, we foresee that almost all verbs to be merged will require detailed manual procedures.

As sketched out in Section 4, an encoding tool is foreseen to facilitate this task of establishing links by aligning SynUs and synsets of a given lemma, leaving for the lexicographer to specify the merge types.

Acording to our estimations, a project size of 1 1/2 man years should be realistic for completing a full merge, leaving approx. four man months for the task of establishing automatic links between unproblematic nouns and for the development of the encoding tool for the more complex cases.

With start in 2008, a first phase of the merge of STO and DanNet is included as a work package in the CLARIN-DK project (Common Language Resources and Technology Infrastructure) supported by the Danish Ministry of Research. CLARIN-DK is connected to the corresponding EU CLARIN project and shares its goals – only at a national level. The CLARIN projects are committed to establish an integrated and interoperable research infrastructure of language resources and its technology. In other words, the goal is to lift the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and thereby enabling eHumanities at a larger scale (cf. www.clarin.eu).

6. References

DDO = Hjorth, E., Kristensen, K. et al. (eds.). Den Danske Ordbog 1-6 (‘The Danish Dictionary 1-6’). Gyldendal & Society for Danish Language and Literature (2003-2005).

Asmussen, L. Pedersen, B.S. & Trap-Jensen, L. (2007). DanNet: From Dictionary to WordNet. Kunze, C., Lemnitzer, L. & Osswald, R. (eds.) *GLDV-2007 Workshop on Lexical-Semantic and Ontological Resources* 1--11. Universität Tübingen, Germany.

Braasch A., Olsen S.A.: (2004). STO: A Danish Lexicon Resource - Ready for Applications. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation* s. 1079--1083. Lisboa, Portugal 2004.

Lenci, A. et al. (2000). SIMPLE – A General Framework for the Development of Multilingual Lexicons. In T. Fontenelle (ed.) *International Journal of Lexicography Vol 13,* pp. 249--263. Oxford University Press

Pedersen, B. & P. Paggio (2004). The Danish SIMPLE Lexicon and its Application in Content-based Querying In *Nordic Journal of Linguistics Vol 27:1* p.97--127.

Pedersen, B.S. & S. Nimb. (2008). Event Hierarchies in DanNet. In: A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, & P. Vossen (eds) *Proceedings of Global WordNet Conference*, Szeged, Hungary, pp. 339--349.

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

Roventini, A. & N. Ruimy (2006). Linking and Harmonizing Different Lexical Resources: a Comparison of Verbal Entries in ItalWordNet and PAROLE-SIMPLE-CLIPS. In: *Proceedings of the Third International WordNet Conference* pp.251--259, Korea.

Ruimy, N. O. Corazzari, E. Gola, A. Spanu, N. Calzolari, A. Zampolli (1998). ‘The European LE-PAROLE Project: The Italian Syntactic Lexicon’, in: *First International Conference on Language Resources & Evaluation*, Granada, Spain.

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

Vossen, P., K. Hofmann, M. de Rijke, E.T.K. Sang, & K. Deschacht (2007). The Cornetto Database: Architecture and User-Scenarios. In *DIR 2007,* pp. 89--96. Leuven, Belgium.