The Swedish WordNet Project

Åke Viberg
Uppsala University, Department of linguistics
Box 527
SE-751 20 Uppsala/Sweden
Sweden
Ake.Viberg@ling.uu.se

Co-authors: Kerstin Lindmark & Ann Lindvall, Lund University;
Ingmarie Mellenius, Umeå University;
Kerstin.Lindmark@ling.lu.se, Ann.Lindvall@ling.lu.se, Ingmarie.Mellenius@nord.umu.se

Abstract

The Swedish WordNet project aims at building a Swedish version of the EuroWordNet lexical database. The article accounts for some of the problems specific to the building of a Swedish net. As an illustration, three examples are discussed: the coding of words denoting human beings, musical instruments and motors, and, finally, verbs denoting change.

1 Introduction

This paper reports on work in progress on a Swedish version of EuroWordNet. The project Swedish WordNet is funded for three years 2000-2002 by the Swedish Research Council for the Humanities and Social Sciences. Swedish WordNet is structured according to the principles of the original Princeton WordNet [Fellbaum 1998] and in particular to its sequel EuroWordNet (EWN) [Vossen 1999]. The basic unit in the wordnets is a synset, a set of synonyms which represent a certain meaning. The synsets are related according to a number of semantic relations such as hyponymy, meronymy and antonymy. At the end of 2001, around 15 000 synsets have been coded in the Swedish WordNet but the aim is to reach 40 000-50 000 synsets (primarily Nouns and Verbs). The coding also includes the linking of the Swedish synsets to the interlingual index provided by EuroWordNet which makes it possible to link synsets across all the languages for which such nets exist. Basically, links are provided from synsets in the other languages to the closest synsets in version 1.5 of the Princeton WordNet.

Various methods have been used to construct the existing versions of EWN. The two main strategies have been the expand approach where WordNet 1.5 synsets have been translated to another language and the merge approach where a separate net has been constructed and then mapped to WordNet 1.5. [Vossen et al. 1999]. The Spanish wordnet followed the first approach whereas the Dutch and Italian wordnets followed the second approach, which is also the approach being followed by the Swedish WordNet project.

In the EWN, synsets can be linked to a Top Concept ontology consisting of a hierarchy of 63 language-independent concepts such as (to pick a few examples) Plant, Human, Animal, Vehicle, Furniture, Building for concrete nouns (1st Order Entities) and Communication, Existence, Location, Possession for abstract nouns and verbs (2nd Order Entities). The coding of the Swedish net to a great extent has proceeded top concept by top concept in order to uncover the semantic relations between Swedish words as efficiently as possible. For 2nd Order Entities, the top concepts to a rather great extent coincide with semantic fields used in earlier work on Swedish verbs [e.g. Viberg 1981, 1999]. Since the general architecture of the lexical database is very well documented (see [Web sites] for WN and EWN), the rest of this presentation will concentrate on the lexical semantic structure of Swedish, problems with its coding and the relations to the coding in wordnets for other languages

2 The Treatment of Words Denoting Human Beings in the Wordnets

One of the most common functions of words in languages in general is to denote individuals. In ordinary speech, this is mostly achieved by the use of proper names and personal pronouns. Still, in a language such as Swedish a fairly large proportion of the lexicalized words are nouns denoting persons. An examination of the word list published by the Swedish Academy [SAOL 1986] reveals that around 7 500 words, or more than 6% of the 120 000 entries, are nouns of this type. As SAOL in all likelihood gives a well-balanced picture of the Swedish lexicon, a lexical resource such as the Swedish WordNet, that is limited to nouns and verbs, should contain an even larger proportion of nouns denoting individual human beings.

Is it likely that other languages contain lexicalized words denoting individuals to the same extent as Swedish? This is a question which is difficult to answer. However, to some extent it is possible to study the proportion of words of this kind in the existing wordnets for different languages. In the EWN 1 project (the Dutch, English, Italian and Spanish wordnets), 1144 base concepts were chosen to form a set of so-called common base concepts (CBCs), which wordnet builders of every language should take care to include, and which also can be taken as a starting-point for new wordnets. Among the CBCs, 106, or 9,3%, are located under the top concept 'human'. This amount comprises 43 concepts that are cross-classified under 'human' and 'group' (e.g., "administration 3", "company 1"), some of which are also classified as 'function' (e.g. "company 2", "institute 1"). Only one BC is 'part + human' ("department 1"); the remaining 62 base concepts classified as 'human' are applicable to individual human beings. This means that 5,8% of the CBCs are concepts referring to persons.

Vossen et al. [1998, p. 13] report on the number of synsets that have been classified under the different top concepts in EWN 1. As a proportion of the total wordnets according to statistics in Vossen [1998], the numbers they report imply that in WordNet 1.5 12,6% of the synsets are found under the top concept 'human'; in the Dutch wordnet 14,5%, in the Spanish wordnet 33%, and in the Italian wordnet 11% of all synsets that are classified as 'human'. Note that the synsets under 'human' are cross-classified with other categories (such as 'group'), so that synsets used about individual persons form a subset of this set. A means of getting closer to this subset is to use statistics in Vossen et al. [1998, p. 16] on Dutch, Spanish and Italian nouns clustered over lexicographer's file codes used for WordNet 1.5.

This system also uses cross-classification, but the number of synsets under 'noun.person' is smaller than the corresponding number under the top concept 'human'. In WordNet 1.5, we find 10,2% of the synsets in the wordnet under 'noun.human'; in Dutch, 12%, in Spanish, 29%, and in Italian, 9,3%.

How does the wordnet builder go about the task to organize several thousand words under the common hyperonym 'person'? In broad outlines, this can be achieved either through shallow hierarchies, yielding flat structures, with many direct hyponyms under each node, or through deep hierarchies, where each node has fewer direct hyponyms but where the hyponym chains are longer. The Dutch wordnet structures words for persons more along the first line, so that *mens* gets 572 direct hyponyms. The Geman wordnet, on the other hand, seems to put more emphasis on structuring and building hierarchies: there are only 12 direct hyponyms under *Mensch.* Some of the hyponyms to *Mensch* are base concepts in the German net, although they are not in the set of CBCs: this is the case of e.g. *Charakterbeschaffener*, under which many concepts are found: concepts that in 1.5 are sorted under 'good person', 'bad person', etc. The case of *Charakterbeschaffener* illustrates a problem in the structuring of words for persons. According to EWN principles, it is possible to have a hyponym relation that goes between words from different word classes, e.g. nouns and verbs, but it is not as easy to permit hyponymy between nouns that are 1stOrderEntities (i.e. concrete nouns) and nouns that are 2ndOrderEntities (i.e. abstract nouns).

In the Swedish WordNet, there are around 20 direct hyponyms to *människa.* Although a pronounced goal of the EuroWordNet initiative is not to include levels that are not lexicalized in the language, it is obvious that many nouns for persons are grouped together not because they are co-hyponyms of the same synset, but rather because they are instances of the same phenomenon. An example of this is nouns like 'angler', 'philatelist' etc. which can be grouped together by the use of words like 'interest', 'pastime', 'hobby'. However, in cases like this, when no proper hyperonym is lexicalized in the language, a noun phrase is used in the Swedish net. The existence of lexical gaps is thus revealed by the existence of phrases in the wordnet. If levels of organisation that are not represented by words in the language were not to be included in the wordnet, it would be extremely difficult to detect lexical gaps.

3 The Treatment of some Domain-Specific Concepts

Even for domain-specific applications, the wordnet format offers interesting possibilities, for example for translators. To this end, however, a high degree of precision is required.

3.1 Musical Instruments

Considering the amount of different principles that can be used for categorising musical instruments, the classification is not obvious, and many such attempts have been made by musicologists. For the work with musical instruments within Swedish WordNet, a "light" version of such a classification was made available [Edlund 1976]. This produced a very well-structured subnet with well-defined categories, the first hyponyms being idiophones, aerophones, electrophones, membranophones and chordophones. In WordNet 1.5, there are 10 direct hyponyms, *bass* being one of these.

Whereas WordNet 1.5 is rather a kind of inventory of common words in the field, with many specific instruments occurring directly under *musical instrument*, the Swedish version can be used for investigating how different instruments actually relate to each other. A problem from a wordnet perspective may be that some synsets consist of phrases rather than words, in some cases rather technical descriptions. In this domain, however, these phrases are commonly used, and are thus not artificial lexical items constructed for wordnet purposes or labels used for distinguishing different categorisation criteria. Where there is a scientific taxonomy which is transparent for non-scientists, this should be used.

3.2 Motors

Motors can be classified according to several different criteria, such as input energy, energy medium, inventor, parts, working principle [TNC 1984]. Also "purpose" or "application" would be a relevant criterion. In WordNet 1.5, no distinction between these different criteria is made. A small part of the structure is shown in Figure 1.

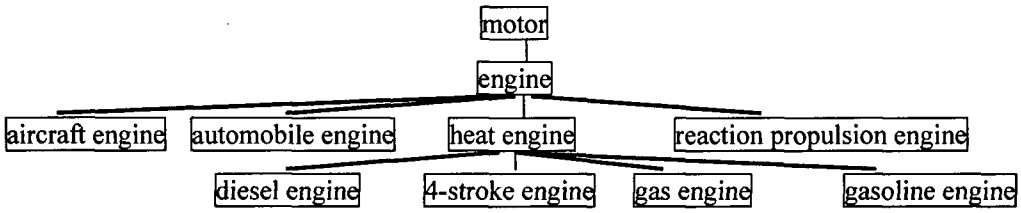

Figure 1. Hyponymy relations of " motor" in WordNet 1.5 (selection).

Thus, *automobile engine* which refers to application is coordinate with *heat engine*, which relates to input energy medium, and there is no direct link between e.g. *gasoline engine*, which is the most common kind of motor that is used in a car, and *automobile engine*. In EWN, this problem has been discussed [Vossen & Bloksma 1998], but the solutions suggested seem not to have been implemented. Within Swedish WordNet, different ways of handling the multidimensionality problem have been tried. One solution would be to place "application" or "function" along one axis and e.g. "working principle" along another and assign every synset a hyperonym from each axis. An alternative solution is to try to follow one axis and make the distinctions criterion by criterion as far as possible and then have hyponyms according to different criteria on the same level, just like Princeton WordNet, but grouping them and assigning them the labels "conjunctive" and "disjunctive", so that *automobile engine* is conjunctive with *gasoline engine* and disjunctive with, e.g., *airplane engine*.

4 The Treatment of Verbs, Especially Verbs Denoting Change

Verbs and abstract nouns – 2ndOrderEntities – in the EWN are linked to two types of top concepts, namely SituationTypes (Dynamic–Static, Bounded–Unbounded etc) and SituationComponents (Cause, Condition, Communication, Possession etc.). The Swedish WordNet follows this analysis but in addition to this makes a systematic distinction between transitive and intransitive verbs.

A special category of verbs is the one called Verbs of Change in Princeton WordNet. In the EWN, these are assigned varying SituationComponents, such as Condition (*worsen, improve*), Physical (*redden, thicken, widen, enlarge*) and Quantity (*lessen, increase, decrease*). The major hyperonym in English is *change*. The 9 senses of this verb in the English WordNet correspond to meaning distinctions in Swedish. However, while the different English senses are defined and specified by the other members of the synset (e.g. *alter, change*, opposed to mere *change*), the corresponding meanings are expressed by different lexemes in Swedish: *ändra / förändra / byta / växla* etc. There is no Swedish hyperonym that occurs in all these senses. Swedish has two hyperonyms, according to how fundamental the change is: *ändra* and *förändra*, respectively. Furthermore, intransitive and transitive verbs in general are distinguished through systematic morphological means [Lindvall 2002]. Dutch and German that are also Germanic languages have few polysemous verbs denoting change but instead have separate verbs for each sense. The Dutch version has *veranderen / wijzigen / wisselen*, the German one has *ändern / verändern / wandeln / wechseln* etc. On the other extreme, the French version has 20 senses of the verb *changer* but few synonyms. As shown in Figure 2, the French *changer* and to a lesser extent the English *change* cover a wider range of senses, whereas the corresponding senses are distributed across a range of separate verbs in Swedish, Dutch and German.

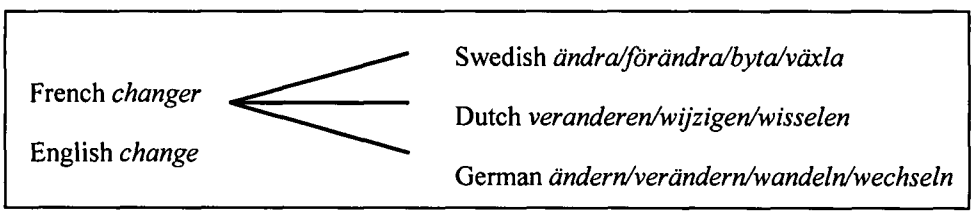

Figure 2. Contrastive relationships between major verbs denoting 'change'

Another difference among the language-specific wordnets is that verbs are categorized under different hyperonyms. Many verbs denoting production are treated by the Swedish WordNet as troponyms of the hyperonym *framställa, göra* 'create, make', e.g. *blåsa (ett glas)* 'blow (a glass)', *trycka (en bok)* 'print (a book)' etc. This is justified by the fact that the glass comes into existence as a result of the blowing. In the English analysis, such verbs are troponyms under *alter, change*, focussing the activity prior to the result.

Swedish verbs like *uppfostra (ett barn)* 'socialize (a child)' and *utbilda* 'educate' are regarded as troponyms of *interagera* 'interact' based on the conception of such activities as mutual and interactive. In the German WordNet, the corresponding verbs are treated as hyponyms of *lehren* 'teach' and further of *agieren, handeln* 'act', whereas they are treated as troponyms of *improve* in the English WordNet and further of *alter, change*, focussing the change from one state to another.

5 Conclusion

In this short presentation, it has only been possible to give restricted examples of the decisions that have to be made in the coding. In any case, what has been said should suffice

to show that many interesting observations can be made with respect to contrastive relationships between the languages for which wordnets are now available. However, much work remains to be done in order to harmonize the coding across languages, since in many cases it is difficult to determine whether there is a genuine contrast in semantic structure or only a spurious difference due to differences in methodology or available resources.

References

[Edlund 1976] Kompendium över europeiska musikinstrument av Bengt Edlund omarbetad och utvidgad av Claes af Geijerstam och Lars Hallgren. Uppsala 1976.

[Fellbaum 1998] Fellbaum, C. (ed.), 1998. *WordNet. An Electronical Lexical Database*. The MIT Press, Cambridge, Mass., US.

[Lindvall 2002] Lindvall, A., 2002. Verbs expressing change – a model, pp.1-20. To appear.

[SAOL 1986] *Svenska akademiens ordlista över svenska språket*. 1986. 11th ed. Norstedts förlag, Stockholm.

[TNC 1984] Energiordlista. *Tekniska nomenklaturcentralens publikationer, nr 81*, 1984. Stockholm.

[Viberg 1981] Viberg, Å., 1981. *Studier i kontrastiv lexikologi*, Ph.D-thesis, Department of Linguistics, Stockholm University, Sweden.

[Viberg 1999] Viberg, Å., 1999. Polysemy and differentiation in the lexicon. Verbs of physical contact in Swedish, in , J. Allwood & P. Gärdenfors (eds.) *Cognitive Semantics. Meaning and Cognition*, pp. 87-129. Benjamins, Amsterdam.

[Vossen 1998] Vossen, P. (ed.), 1998. The Final Wordnets for Dutch, Spanish, Italian and the English Addition, EuroWordNet (LE-4003) Deliverable D032D033, University of Amsterdam.

[Vossen 1999] Vossen, P. (ed.), 1999. *EuroWordNet: a multilingual database with lexical semantic networks for European languages*. Kluwer, Dordrecht.

[Vossen & Bloksma 1998] Vossen P. & Bloksma, L., 1998. Categories and classifications in EuroWordNet, in: A. Rubio, N. Gallardo, R. Catro and A. Tejada (ed) *Proceedings of First International Conference on Language Resources and Evaluation, Granada, 28-30 May 1998*, pp. 399-408.

[Vossen et al. 1998] Vossen, P., Climent, S., Marti, M. A., Taule, M., Gonzalo, J., Chugur, I., Verdejo, F., Escudero, G. Rigau, G., Rodriguez, H., Alonge, A., Bertagna, F., Marinelli, R., Roventini, A., Tarasi, L. 1998. Comparison of the Final Wordnets Dutch, Spanish and Italian, EuroWordNet (LE-4003) Deliverable D029D030, University of Amsterdam.

[Vossen et al. 1999] Vossen, P., Bloksma, L. & Boersma, P. 1999. The Dutch Wordnet. Deliverable D032, D033.

Web sites

Detailed information about the wordnets are available on the two following web sites:
(Princeton) WordNet: http://www.cogsci.princeton.edu/~wn
EuroWordNet: http://www.hum.uva.nl/~ewn. All the deliverables referred to above can be found here.