The Cornetto database: Semantic issues in linking lexical units and synsets

Hennie van der Vliet, Isa Maks, Piek Vossen and Roxane Segers
Faculteit der Letteren, Vrije Universiteit Amsterdam

Cornetto is a lexical semantic database that combines the Dutch Wordnet (Vossen 1998) and the Referentie Bestand Nederlands (Van der Vliet 2007). The Dutch Wordnet (DWN) is similar to the Princeton Wordnet for English (Fellbaum 1998), and the Referentie Bestand Nederlands (RBN) includes frame-like information as in FrameNet (Fillmore, Baker, Sato 2004) as well as information on the combinatorical behaviour of word meanings. The combination of the lexical resources has resulted in a rich relational database that may improve natural language processing technologies.
An important aspect of combining the resources is the alignment of the lexical units (LU's) and the synsets. Automatic alignment of RBN and DWN resulted in an initial version of the Cornetto database. This version has been further extended both automatically and manually. The resulting data structure is stored in a database that keeps separate collections for LU's (mainly derived from RBN), synsets (derived from DWN) and, in addition, a formal ontology (SUMO/MILO, see Niles and Pease 2001). These 3 semantic resources represent different viewpoints and layers of linguistic and conceptual information. The resulting resource is freely available for research in the form of an XML database.
In this contribution, we will concentrate on the semantic information in Cornetto. We will discuss the differences in the perspective on semantics in the LU's and synsets and we will give a brief overview of the differences with regard to semantic information. The merging of the two resources resulted in very rich semantic database. However, combining lexica with different perspectives on semantics causes specific problems in the alignment of LU's and synsets and leads to findings that shed light on the organization of meaning in the lexicon.

1. Introduction

Cornetto is a lexical semantic database that combines the Dutch Wordnet (Vossen 1998) and the Referentie Bestand Nederlands (Van der Vliet 2007). The Dutch Wordnet (DWN) is similar to the Princeton Wordnet for English (Fellbaum 1998), and the Referentie Bestand Nederlands (RBN) includes frame-like information as in FrameNet (Fillmore, Baker, Sato 2004) as well as information on the combinatoric behaviour of word meanings. The combination of the lexical resources has resulted in a rich relational database that may improve natural language processing technologies, such as word sense-disambiguation, and language-generation systems.

An important aspect of combining the resources is the alignment of the lexical units (LU's) and the synsets. Automatic alignment of RBN and DWN resulted in an initial version of the Cornetto database. This version has been further extended both automatically and manually. The resulting data structure is stored in a database that keeps separate collections for LU's (mainly derived from RBN), synsets (derived from DWN) and, in addition, a formal ontology (SUMO/MILO, see Niles and Pease 2001). These 3 semantic resources represent different viewpoints and layers of linguistic and conceptual information. The resulting resource is freely available for research in the form of an XML database.

In this contribution, we will concentrate on the semantic information in Cornetto. We will discuss the differences in the perspective on semantics in the LU's and synsets and we will give a brief overview of the differences with regard to semantic information. Combining different perspectives causes specific problems in the alignment of LU's and synsets, but the merging of the detailed semantic information from both resources will also lead to very rich semantic descriptions in the resulting Cornetto database.

In section 2, we first give an overview of the structure of the database with emphasis on the semantics. Then the lexical organization of the RBN and Dutch WordNet will be discussed. In

section 3, we discuss the problems of alignment and in section 4. the benefits of merging two lexical databases with different organizational principles. In the last section we will summarize and comment on some findings.

2. Semantics in the Cornetto database

The Cornetto database (CDB) consists of 3 main data collections,: the RBN, the DWN and the Cornetto ontology (COON), a collection of terms and axioms.

The RBN contains orthographic, morphological and pragmatic information and is especially rich in syntax and semantics. The LU's in the RBN are associations of form units (word forms, roughly speaking) and meanings. As such, they correspond to word senses in the lexical semantic tradition (Cruse 1986).

DWN is a organised on the notion of Synsets. Synsets are sets of synonyms. They form concepts in a relational model of meaning, as defined by Miller and Fellbaum (Miller et.al. 1990, Fellbaum 1998). Synsets are conceptual units, but strictly related to the lexicalization pattern of a language and defined by lexical semantic relations.[1] Typically in Wordnet, information is provided for the synset as a whole and not for the individual word meanings. For example, in Wordnet the synset has a single gloss but the different LU's in RBN each have their own definition. It follows that from a Wordnet point of view, the definitions of LU's that belong to the same synset should be semantically compatible or synonymous.

A third layer of meaning, but outside of the lexicon, is the ontology. In the ontology meaning is defined independently of language but according to the principles of logic. In this contribution, we will not go into the ontology but we will concentrate on the semantics in the lexicon.

Next to the three data collections, a table of so-called Cornetto Identifiers (CIDs) is provided. The CID's tie together the separate collections of LU's and Synsets and as such they are just administrative records. The lexical data are stored in the collection of LU's and in the collection of synsets[2].

Table 1 gives an overview of the most important syntactic and semantic features in the LU of the first meaning of lopen (*to walk*). Orthographic, pragmatic and morphologic information, as well as twelve additional examples are left out, because they are less relevant here.

[verb] lopen 1

Syntax:
trans: intr separ: onsch class: main peraux: h/z valency: di reflexiv: nrefl subject: pers
complementation: nil, pp [..]

Semantics:
type: action
caseframe: mvmt2 (caserole: agent selrestrole: agentanimate , caserole: soudirpath selrestrole: soudirpanselres)
definition: zich stappend voortbewegen (*moving step by step*)

[1] For Cornetto, the semantic relations from EuroWordNet are taken as a starting point (Vossen 1998).

[2] For more details on the architecture of the Cornetto Database see Vossen et al. 2008.

Examples:	
naar huis lopen (*walk home*)	type: free
trappen lopen (*go up and down the stairs*)	type: fixed subtype: lexcol
je de benen uit je lijf lopen (*run like mad*)	type: fixed subtype: idiom
achter iemand aan lopen (*run after someone*, also fig.)	type: fixed subtype: idiom
in [de steun/WW/...] lopen (*be on the dole*)	type: fixed subtype: idiom
loop naar de maan! (*get stuffed*)	type: fixed subtype: pragma

Table 1

As part of the syntactic information, all possible complementation patterns are explicitly listed. This meaning of *lopen* typically will be used without a complement (the *nil*-option) or with a prepositional phrase. The corresponding semantic properties are worked out in a caseframe. The LU *lopen 1* is described as an action verb of the type *mvmnt 2* and this type is associated with two case roles, *agent* and *source/direction/path* for the prepositional phrase. The information on complementation patterns and the caseframe is reflected in the examples. The free examples are more or less productive, whereas the fixed examples are restricted in a syntactic, semantic or pragmatic way. As a rule, all complementation patterns are illustrated by free or fixed examples. In addition there are semantic and syntactic collocations, pragmatic formulae and idiomatic expressions.

In table 2 we give a brief overview of the information of the corresponding synset of *lopen 1*in Dutch Wordnet.

Synonyms: gaan:13/r_v-2848, lopen:1/r_v-4435, treden:5/d_v-294705

PoS specific: VERB_INTRANSITIVE
Definition: zich met de benen voortbewegen
SUMO: (=, , Walking)

-->> [HAS_HYPERONYM] voortbewegen:2 (*move on*)
<<-- [HAS_SUBEVENT] uitlopen:15 (*walk out of*)
-->> [HAS_XPOS_HYPONYM] loop:3(*run, flight*)
-->> [INVOLVED_AGENT] loper:4 (*walker*)
<<-- [INVOLVED_AGENT] voetganger:1 (*pedestrian*)
<<-- [INVOLVED_INSTRUMENT] looprek:1, loophek:1 (*walking frame*)
-->> [INVOLVED_INSTRUMENT] poot:1, pootje:2 (*paw*)
-->> [INVOLVED_INSTRUMENT] onderdaan:2, onderdanen:1, stelt:2, been:1, poot:10 (*leg*)
<<-- [INVOLVED_LOCATION] pad:1 (*path*)
<<-- [INVOLVED_LOCATION] voetpad:1 (*footpath*)
<<-- [IS_CAUSED_BY] kreupelheid: 1 (*lameness*)
<<-- [XPOS_NEAR_SYNONYM] lopend:1 (*walking, going*)

Equivalence relations: [EQ_SYNONYM] /ENG20-0184

Table 2

As you can see *lopen 1* is in a synset with *gaan 13* and *treden 5*. The verb is marked as intransitive, there is a definition, a link to the SUMO-ontology and a link to the Princeton English Wordnet. The synsets are conceptually characterized by their relations. As table 2 shows, a rich semantic network is created by situating *lopen 1* in a synset and by relating this synset by various relations to other synsets.

In creating the Cornetto database the alignment of the LU's and the synsets is a crucial step. In this example there is a match for *lopen 1* as a LU and as a synset, but this process of

alignment is not always straightforward. In the next section we will go into some details on the process of aligning.

3. Combining Lexical Units and synsets: the problems

To create the initial database, we performed an automatic alignment of word meanings (see Vossen et al. 2006 for more details). All possible mappings were generated for the system with confidentiality scores. The evaluation showed that the automatic alignment the form units with a single meaning in RBN and DWN and a lot of the bisemes are mostly correct. The next step was therefore the manual aligning of low scoring meanings and meanings without links. If no links were found, often this was caused by the fact that DWN has a larger macro structure then the RBN. In these cases we created a new LU matching the DWN-word in the synset. In some cases the DWN-word was very marginal (old fashioned or strictly regional), and was removed.

As a next step in alignment, we identified four groups of problematic cases: frequent polysemous verbs and nouns, nouns with a semantic shift, adjectives and multi word units. We will focus here on the semantic problems we faced with the polysemous words and the shifts.

It is obvious that polysemous words are hard to align automatically. At the same time they are often very frequent and that is why they are particularly in need of manual editing. The following cases are typical:

- a meaning is missing as a LU or as a synset => LU or synset is added
- a LU corresponds to two or more synsets =>LU is split up or a synset is removed
- a synset corresponds to an idiomatical combination in a LU => LU is added.

In large and complex entries sometimes these problems can be found simultaneously.
Most of these alignment problems are caused by the differences in basic assumptions underlying the meaning description in RBN and DWN. The RBN aims to deal with semantics in a systematic and efficient way, the DWN on the other hand is much more fine grained on the meaning description.

As an illustration we present a case where a single LU corresponds to two synsets, typically occurring when a LU has a literal and a metaphorical meaning and in DWN the metaphorical meaning corresponds to different synsets..An example is the verb *brouwen.* This word form is associated with two LU's (to brew beer and to burr, to pronounce the r-sound in an a-typical way) and with four synsets (in addition to the LU's: *preparing a meal* and *making plans*). Both additional synsets are cases of metaphorical meaning extension to the *beer brewing* meaning, in which *brouwen* has an additional association of preparing, making or inventing something in a somewhat obscure way. The meanings in the two additional synsets can be seen as more or less creative utterances based on a general metaphor[3]. The point is that the presence of these two DWN meanings seems to be motivated on the basis of the already existing synsets; for both meanings, *brouwen* is a possible lexicalization. In the RBN on the other hand, the meanings are described with the form unit as a point of reference. The metaphorical meanings did not lead to separate LU's, but they are recognized as a single

[3] The background of this metaphor is that *brouwen* is the standard verb used for the preparation of obscure mixtures by the dark powers of evil, for instance in fairy tales. This may well be the most frequent meaning of the verb, but it was missed by RBN and DWN. We added this meaning in the Cornetto database.

metaphor and represented by an idiomatic expression *er niets van brouwen* (to make a complete mess of something). How should the alignment be performed? Creating new LU's for the additional synsets does not seem to be a proper solution. We will come back to this problem in the last section.

The nouns with semantic shifts are characteristic for the way the RBN is dealing with semantics. One of the strategies to deal with word meaning in a systematic way is the use of systematic meaning shifts, like the shift from Process to Action in verbs and from Dynamic to Non Dynamic in nouns. For example the noun *bekendmaking* (*announcement*) is represented in a single LU, with a meaning shift from a dynamic (the announcing) to a non-dynamic (the anouncememt) reading. In DWN however, these are separate word meanings. As a consequence, in order to align the LU's and the synsets, the LU's with semantic shifts are candidates for splitting up.

All in all, the automatic alignment was quite successful. In the more polysemous entries however, we experienced that even manual alignement could be a very hard job. This is the direct result of the principles underlying the semantic organisation of the resources. Therefore it is inevitable, that the alignment of LU's and synsets leads to a unified view on semantic description, which is at the cost of some of the characteristic information in the resources.

4. Combining the Lexical Units and Synsets: the benefits

As table 1. and 2. show, the semantic information from the LU's and the synsets are complementary. The LU's for the verbs in the RBN are very explicit on complementation, case frames and combinatorics (the syntagmatic relations), while the corresponding synsets build a conceptual network based on a rich set of lexical relations (the paradigmatic relations). In addition to this, the RBN also offers very detailed syntactic, pragmatic and morphological information.

As an example of combining the examples in the LU's with the semantic network we will discuss the domain of *drinks*. In Dutch, the preparation of drinks is usually referred to by the general verb *maken* (*to prepare*). However, in the case of *koffie* (*coffee*) and *thee* (*tea*), another specific verb is used: *zetten*. One typically uses the phrases *koffie zetten* and *thee zetten* (*to make coffee* and *to make tea*) but in case you prefer lemonade, you should use the standard phrase *limonade maken* (*to make lemonade*) in Dutch. This example illustrates that conceptual combinations and constraints that are encoded in the synsets, do not always explain the proper (and often most frequent!) way of phrasing relations. In these cases the combinatorical information in the LU's will help the user to find the correct phrasings.

5. Conclusion and discussion

The RBN provides detailed information on morphological, semantic, syntactic and combinatorical information on the LU's, the synsets place these LU's within a rich semantic network. The result of combining these resources is a very rich and multi-functional lexical database. However, the two lexical resources are based on different meaning perspectives and the alignment of LU's and synsets leads to a unified view on meaning. The example of *brouwen* shows what the problem is: a synset-meaning may be lexicalized in specific ways, but seen from the perspective of LU's these lexicalizations are not always the most natural meanings of a word form. On the other hand, semantic information in the LU combinations cannot always be encoded straightforward in synsets. This leads to problems in the alignment,

as in the example of *brouwen*. The combinatorics, especially the fixed combinations, may partly bridge the gap. By representing fixed combinations as multi word units and linking them to synsets, this problem can be solved in a very natural and straightforward way. From the LU point of view, synsets make it possible to describe the meaning of a frequent, yet non-compositional combination. From a synset perspective, the LU's are very helpful in detecting the idioms that, as multi word units, can function as part of a sysnet. As a result, the two synsets for *brouwen*, *preparing a meal* and *making plans*, should not lead to additional LU's, but to the additional multi word units *een maaltijd brouwen* and *een plan brouwen* in the same LU. These findings are of interest for the organization of meaning in the lexicon since they clearly reveal the importance of combinatorical constructions for word meaning.

References

Cruse, D. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Fellbaum, C. (ed.; 1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, C.; Vossen, P. (2007). 'Connecting the Universal to the Specific: Towards the Global Grid'. In *Proceedings of the First International Workshop on Intercultural Communication*. Reprinted in: 'Intercultural Collaboration: First International Workshop'. In Ishida, T. I, Toru, Fussell, S.R. and Vossen, P. T. J. M. (ed.). *Lecture Notes in Computer Science*. New York: Springer. Vol. 4568. 1-16.

Fillmore, C.; Baker, C.; Sato, H. (2004). 'Framenet as a 'net''. In *Proceedings of Language Resources and Evaluation Conference (LREC 04)*. Lisbon: ELRA. Vol. 4, 1091-1094.

Maks, I.; Martin, W.; Meerseman, H. de (1999). *RBN Manual*. Amsterdam: Vrije Universiteit

Maks, I.; Vossen, P.; Segers, R.; Vliet, H. van der (2008). 'Adjectives in the Dutch semantic lexical database CORNETTO'. In *Proceedings of LREC-2008*. Marrakech.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K. (1990). 'Introduction to WordNet: An On-line lexical Database'. In *International Journal of Lexicography* 3 (4). 235-244.

Niles, I.; Pease, A. (2001). 'Towards a Standard Upper Ontology'. In *Proceedings of FOIS 2001*. Maine: Ogunquit. 2-9.

Vossen, P. (ed.; 1998). *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Dordrecht: Kluwer.

Vossen, P.; Maks, I.; Segers, R.; Vliet, H. van der; Zutphen, H. van (2008). 'The Cornetto Database: the architecture and alignment issues'. In *Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008*. Hungary: Szeged.

Vliet, H. van der (2007). 'The Referentie Bestand Nederlands as a multi-purpose lexical database'. In *International Journal of Lexicography* 20 (3).