Integrating lexical units, synsets and ontology in the Cornetto Database

Piek Vossen[1, 2], Isa Maks[1], Roxane Segers[1], Hennie van der Vliet[1]

[1] Amsterdam Faculteit der Letteren, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
email: {p.vossen, e.maks, rh.segers, hd.vandervliet}@let.vu.nl
[2] Irion Technologies, Delftechpark 26, 2628 XH, Delft, The Netherlands,
email: piek.vossen @irion.nl

Abstract

Cornetto is a two-year Stevin project (project number STE05039) in which a lexical semantic database is built that combines Wordnet with Framenet-like information for Dutch. The combination of the two lexical resources (the Dutch Wordnet and the Referentie Bestand Nederlands) will result in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. In addition to merging the Dutch lexicons, the database is also mapped to a formal ontology to provide a more solid semantic backbone. Since the database represents different traditions and perspectives of semantic organization, a key issue in the project is the alignment of concepts across the resources. This paper discusses our methodology to first automatically align the word meanings and secondly to manually revise the most critical cases.

1. Introduction

Cornetto is a two-year Stevin project (project number STE05039) in which a lexical semantic database is built that combines Wordnet with Framenet-like information for Dutch. In addition, the database is also mapped to a formal ontology to provide a more solid semantic backbone. The combination of the lexical resources will result in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. The database will be filled with data from the Dutch Wordnet (Vossen 1998) and the Referentie Bestand Nederlands (Maks e.a. 1999). 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) plus much more information on the combinatoric behaviour of word meanings.

An important aspect of combining the resources is the alignment of the semantic structures. In the case of RBN, these are lexical units (LUs) and in the case of DWN these are synsets. Various heuristics have been developed to do an automatic alignment. Following automatic alignment of RBN and DWN, this initial version of the Cornetto database will be further extended both automatically and manually. The resulting data structure is stored in a database that keeps separate collections for lexical units (mainly derived from RBN), synsets (derived from DWN) and a formal ontology SUMO/MILO (Niles and Pease 2003). These 3 semantic resources represent different viewpoints and layers of linguistic, conceptual information. The database is itself set up so that the formal semantic definition of meaning can be tightened for lexical units and synsets by exploiting the semantic framework of the ontology. At the same time, we want to maintain the flexibility to have a wide coverage for a complete lexicon and to encode additional linguistic information. The resulting resource will be made freely available for research in the form of an XML database.

Combining two lexical semantic databases with different organizational principles offers the possibility to study the relations between these perspectives on a large scale. However, it also makes it more difficult to align the two databases and to come to a unified view on the lexical semantic organization and the sense distinctions of the Dutch vocabulary. In this paper, we discuss the alignment issues. In section 2, we first give an overview of the structure of the database. Section 3 describes the approach and results of the automatic alignment. Section 4, discusses the manual work of checking and improving the automatic process. This work mainly involves comparing the LUs from RBN with the synset structure of DWN. Finally, in section 5, we discuss the relation between synsets and the ontology.

2. Architecture of the Database

Figure 1 shows an overview of the different data structures and their relations. The different data can be divided into 3 layers of resources, from top to bottom:

- The RBN and DWN (at the top): the original databases from which the data are derived;
- The Cornetto database (CDB): the ultimate database built;
- External resources: any other resource to which the CDB is linked, such as the Princeton Wordnet, wordnets through the Global Wordnet Association, ontologies, corpora, etc.

The CDB layer consists of 4 major collections that are kept separate: 1) lexical units, 2) synsets, 3) Cornetto identifiers, and 4) possible ontology extensions.
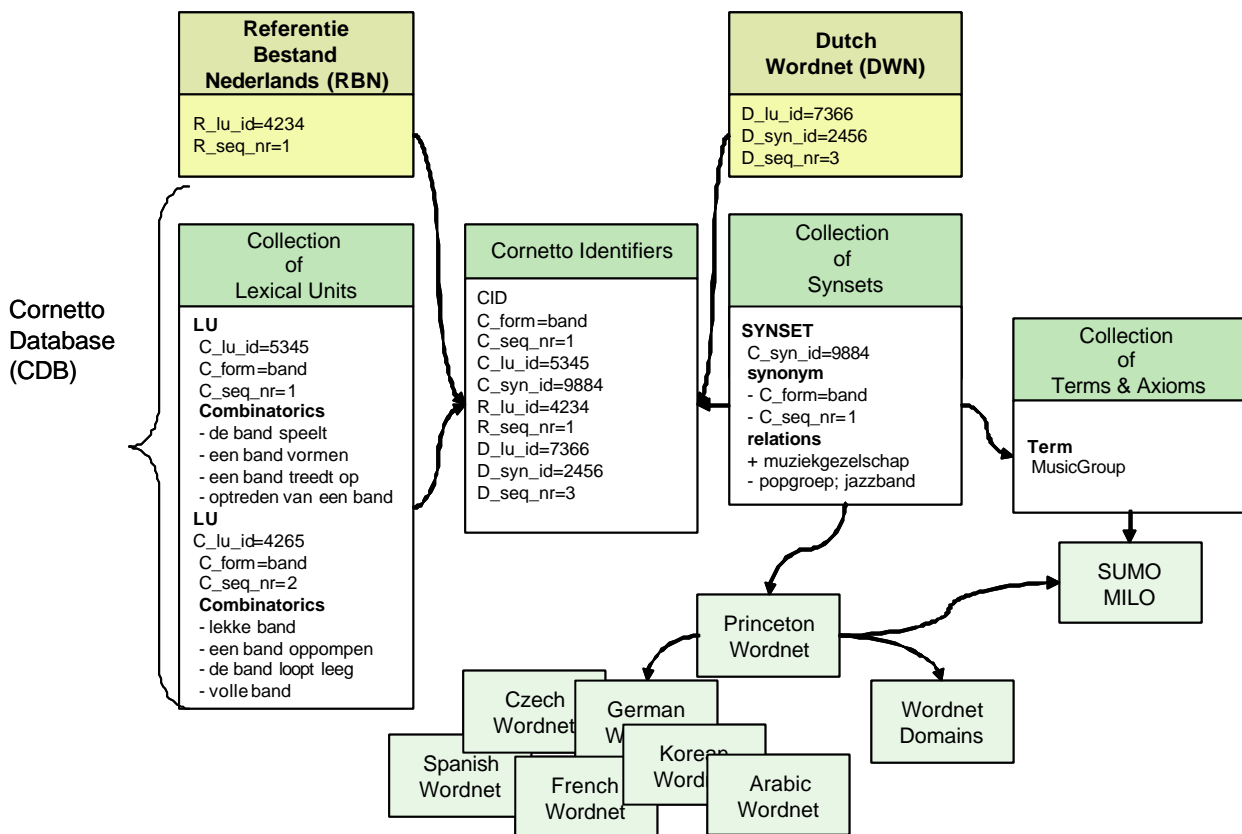
Figure 1: Architecture of the Cornetto databe

The centre of the CDB is formed by the table of the Cornetto Identifiers (CIDs). The CIDs tie together the separate collections of LUs and Synsets but also represent the pointers to the word meaning and synsets in the original databases: RBN and DWN and their mapping relation. In Figure 1, this is shown for the entry "band", which has different meanings in Dutch. The CID record shows one meaning, where the fields that start with "C_" are identifiers for the Cornetto database, fields with "R_" point to RBN and fields with "D_" point to DWN. C_seq_nr=1 indicates that this represents the first meaning in CDB, which corresponds with the first meaning in RBN (R_seq_nr=1) and the third meaning in DWN (D_seq_nr=3). The other identifiers point to lexical units (lu_id) and synsets (syn_id). The structure and content of the identifiers is generated by the alignment process between DWN and RBN.

In the LU collection, we show two lexical units for "band": the first referring to a musical band and the second to an inflated tube or tire, as in a bicycle tire. The Combinatorics fields list typical examples and combination words, but there is much more information on each lexical unit.

The collection of synsets shows a single synset, which here has the lexical unit for musical band as a synonym.

For illustration, we show some semantic relations as in Wordnet. The synset is then related to Princeton Wordnet and through that relation it gets one ore more labels from Wordnet Domains (Magnini & Cavaglià 2000) and the SUMO/MILO mappings.

The Cornetto database provides unique opportunities for innovative NLP applications. The LUs contain combinatoric information and the synsets place these words within a semantic network. The benefits of combining resources in this way are however only possible if the word meanings, representing concepts are properly aligned in the database. This is discussed in the next sections.

3. Aligning automatically RBN with DWN

To create the initial database, the word meanings in the RBN and DWN have been automatically aligned. For example, the word *koffie* (*coffee*) has 2 word meanings in RBN (*drink* and *beans*) and 4 word meanings in DWN (*drink, bush, powder* and *beans*). This can result in 4, 5, or 6 distinct meanings in the Cornetto database depending on the degree of matching across these meanings. This alignment is different from aligning WordNet synsets because RBN is not structured in synsets.

We only consider a possible match between words with

the same orthographic form and the same part-of-speech. The strategies for aligning word meanings are:

1. a word has one meaning and no synonyms in both RBN and DWN
2. a word has one meaning in both RBN and DWN
3. a word has one meaning in RBN and more than one meaning in DWN
4. a word has one meaning in DWN and more in RBN
5. If the broader term (BT) of a set of words is linked, all words which are under that BT in the semantic hierarchy and which have the same form are linked
6. If some narrow term (NT) in the semantic hierarchy is related, siblings of that NT that have the same form are also linked.
7. Word meanings that have a linked domain, are linked
8. Word meanings with definitions in which one in every three words is the same (there must be more than one match) are linked.

Each of these heuristics will result in a score for all possible mappings between word meanings. In the case of *koffie*, we thus will get 8 possible matches with different weights.

The number of links found per strategy is shown in Table 1. To weigh the heuristics, we manually evaluated each heuristics. Of the results of each strategy, a sample was made of 100 records. Each sample was checked by 8 persons (6 staff and 2 students). For each record, the word form, part-of-speech and the definition was shown for both RBN and DWN (taken from VLIS). The testers had to determine whether the definitions described the same meaning of the word or not. The results of the tests were averaged, resulting in a percentage of items which were considered good links. The averages per strategy are shown in Table 1.

The minimal precision is 53.9 and the highest precision is

97.1. Fortunately, the low precision heuristics also have a low recall. On the basis of these results, the strategies were ranked: some were considered very good, some were considered average, and some were considered relatively poor. The ranking factors per strategy are:

- Strategies 1, 2 and 8 get factor 3
- Strategies 5, 6 and 7 get factor 2
- Strategies 3 and 4 get factor 1

A factor 3 means that it counts 3 times as strong as factor 1. It is thus considered to be a better indication of a link than factor 2 and factor 1, where factor 1 is the weakest score. The ranking factor is used to determine the score of a link. The score of the link is determined by the number of strategies that apply and the ranking factor of the strategies.

In total, 136K linking records are stored in the Cornetto database. Within the database, only the highest scoring links are used to connect WordNet meanings to synsets. The other links are also stored in the database and can be selected manually when the automatic link is wrong. There are 58K top-scoring links, representing 41K word meanings. In total 47K different RBN word meanings were linked, and 48K different VLIS/DWN word meanings. 19K word meanings from RBN were not linked, as well as 59K word meanings from VLIS/DWN.

The RBN word meanings are used as a starting point and their numbering and sequence is kept in tact. If there are DWN meanings linked then the RBN meanings are synonyms of the corresponding synsets. If there are no links for an RBN meaning, there is also no matching synset. A synset linkage needs to be created manually, either by finding the matching DWN word meaning, some other synset or by creating a new synset. If there is no match for a DWN word meaning, we create a new dummy lexical unit with the minimal morpho-syntactic information from DWN. The sense number is then sequential to the highest available sense number.

	Conf.	Dev.	Factor	LINKS	
1: 1 RBN & 1 DWN meaning, no synonyms	97.1	4,9	3	9936	8,1%
2: 1 RBN & 1 DWN meaning	88.5	8,6	3	25366	20,8%
3: 1 RBN & >1 DWN meaning	53.9	8,1	1	22892	18,7%
4: >1 RBN & 1 DWN meaning	68.2	17,2	1	1357	1,1%
5: overlapping hyperonym word	85.3	23,3	2	7305	6,0%
6: overlapping hyponyms	74.6	22,1	2	21691	17,7%
7: overlapping domain-clusters	70.2	15,5	2	11008	9,0%
8: overlapping definition words	91.6	7,8	3	22664	18,5%

Table 1. Results for aligning strategies

4. Aligning Manually RBN with DWN

The total number of form units (FUs) in CDB after the automatic alignment is 90.000. About 37.000 occur in both RBN and DWN, 6.000 only occur in RBN and 47.000 only in DWN.

FUs both in RBN and DWN	**37.000**
FUs in RBN only	6.000
FUs in DWN only	47.000
Total number of entries (= FUs) in Cornetto	90.000

Table 2: Form units in the Cornetto database

Most of the shared form units have a single meaning in DWN and RBN. Most of these are also aligned: 22.000 FUs (60%). About 2.200 FUs (6%) with a single meaning in RBN and DWN could not be aligned due to lack of matching information. In the case of biseme FUs, 250 have a direct match across DWN and RBN and 2.500 have at least one non-matching LU or synset.

The monoseme and biseme cases are considered to be reliable. The next manual alignment step consists of editing low-scoring and non existing links between lexical units and synsets. We identified four groups of problematic cases and defined editing guidelines for them which will be presented in the following sections. The manual work involves:

- mapping of RBN and DWN:
- mapping of DWN to English Wordnet 2.0
- mapping of DWN to Wordnet Domain labels
- mapping of DWN to SUMO and MILO

All manual changes are logged and marked. For all automatically derived words and senses, we will extract samples and derive a quality estimate for each of the 4 mappings.

Mapping RBN to DWN involves:
- linking RBN LUs to existing synsets or creating new synsets for unlinked LUs.
- add minimal information to new LUs from DWN
- removing spurious LUs
- removing spurious synsets
- merge or split LUs
- merge or split Synsets

We will discuss this work in more detail in the next four subsections.

4.1 Frequent polysemous verbs and nouns

The low-scoring links within the group of verb synsets and lexical units and within the group of noun synsets and lexical units (following section) are in great deal due to the difference regarding the underlying principles of meaning discrimination which plays an important role in the alignment of synsets and lexical units.

As long as there is a one-to-one mapping from LUs and synsets, the features of the two resources will probably match. Difficulties arise however when the mapping is not one-to-one. Frequent verbs are often very polysemous. The RBN, as the source of the LUs, tries to deal with polysemy in a systematic and efficient way. The synsets, however, are much more detailed on different readings. As a result, in many cases there are more synsets than LUs. In combination with the detailed information on complementation, event structure and lexical relation, this results in interesting editing problems.

A typical example of an economically created LU in combination with a detailed synset is *aflopen* (*to come to an end, to go off (an alarm bell)*, *to flow down, to run down, to slope down, etc.*). Input to the alignment are seven LUs and 13 synsets. Much of the asymmetry was caused by the fact that one of the LUs represents one basic and comprehensive meaning: *to walk to, to walk from , to walk alongside something or someone*. In DWN these are all different meanings, with different synsets. This is the result of describing lexical meaning by synsets; these three readings of *aflopen* obviously have a lot in common, but they match with different synonyms. Aligning the LU's and synsets leads to splitting the LU's and may lead to subtle changes in the complementation patterns, event structure and certainly to adapting and extending the combinatorical information. Sometimes the LUs are more detailed. In that case a synset must be split, which of course gives rise to changes in all related synsets and to new sets of lexical relations. About 1000 most-frequent verbs are manually edited. More details are discussed in Vossen et al 2008.

4.2 Nouns and semantic shifts

The RBN uses a semantic shift label for groups of words that show the same semantic polysemy pattern, represented by a single condensed meaning. DWN explicitly lists these meanings. Because of the difference in approach, the DWN resource will have an extra synset for the meaning that is implied with a shift in the LU. There are about 30 different defined types of shifts that can occur in verbs, adjectives and nouns, like Process ? Action in verbs and Dynamic ? Non-dynamic in nouns.

We expected that the matching of LUs from RBN to synonyms in DWN is more likely to be incorrect for all words labeled with a shift in RBN. We therefore decided to manually verify all the mappings for shifts. The vast majority of 4500 LUs with a semantic shift is found in nouns, on which we have decided to concentrate the manual work.

The editing of these shift cases can be illustrated by the

word *bekendmaking* (*announcement*) that has one LU with a shift in RBN from Dynamic to Non-dynamic. This means that (in Dutch) an *announcement* can be a process and the result of this process. In DWN, we find a synset for each of these aspects, stating that the first one is a subclass of the SUMO term Communicating, and the second one is equivalent to Statement. We can see this as a good argument to split the LU and define the difference in terms of the definition and the semantic relations as is shown in Figure 2:

Dynamic X	announcement
LU resume	'the announcing'
LU combinatorics/example	-
HAS_HYPERONYM	statement (*dyn. in Dutch*)
XPOS_NEAR_ SYN	announcing
SUMO	+ Communicating
Non-dynamic X	announcement
LU resume	'something announced'
LU combinatorics/example	-
HAS_HYPERONYM	message
ROLE_RESULT	announcing
SUMO	+ Statement

Figure 2: example of splitting a LU with a semantic shift

In addition to the nouns with a sift label, we selected all nouns with high-polysemy, i.e. more than 4 meanings (see Table 3). Note that the polysemy after the automatic alignment, is not the real polysemy. If there are no matches, new sense are created.

Nr. of senses after alignment	Nr. words (Nouns)	Nr. senses (Nouns)
10	62	620
9	41	369
8	75	600
7	126	882
6	212	1272
5	389	2235
4	1026	4104
3	2507	7521
2	7293	14586
1	64298	64298

Table 3: Polysemy distribution of nouns

The total amount of manually revised nouns is about 2,000 nouns correlating with about 10,000 senses. This set overlaps with the set of LUs with a shift label.

4.3 Adjectives and fuzzy synsets

A considerable part of the adjectives is not successfully aligned by the automatic alignment procedures. This is especially due to the fact that adjective synsets have few semantic relations lacking hyperonyms and hyponyms. By consequence, the automatic alignment strategies which involve broader and narrower terms are in these cases not applicable.

Another problematic aspect of the adjective synsets is the fact that the automatically formed DWN adjective synsets have not – unlike the noun and verb synsets – been edited and corrected manually before. To be able to deal in a systematic way with these problems, we introduced the use of a semantic classification system for adjectives (Hundschnurser & Splett, Germanet). Details of this work can be found in Maks et al (this volume). About 250 most-frequent adjectives are processed manually.

4.4 Multiword units

Special attention is paid to the encoding and alignment of multiword units. One of the objectives of Cornetto is to introduce part of them, i.e. the fixed combinations into the macrostructure thus making it possible to align them with a synset and via the synset with the ontology. We focus on those combinations which have a reduced semantic (and often syntactic) transparency and a reduced or lack of compositionality.

5. Aligning synsets with ontology terms

5.1 Ontological principles

The ontology is seen as an independent anchoring of concepts to some formal representation that can be used for reasoning. Within the ontology, Terms are defined as disjoint Types, organized in a Type hierarchy where:

- a Type represents a class of entities that share the same essential properties.
- Instances of a Type belong to only a single Type: => disjoint (you cannot be both a *cat* and a *dog*)

Terms can further be combined in a knowledge representation language to form expressions of axioms, e.g. the Knowledge Interchange Format (KIF), based on first order predicate calculus and primitive elements.

Following the OntoClean method (Guarino & Welty 2002a,b), identity criteria can be used to determine the set of disjoint Types. These identity criteria determine the essential properties of entities that are instances of these concepts:

- Rigidity: to what extent are properties of an entity true in all or most worlds? E.g., a *man* is always a *person* but may bear a Role like *student* only temporarily. Thus *manhood* is a rigid property while *studenthood* is non-rigid.
- Essence: which properties of entities are essential? For example, *shape* is an essential property of *vase* but not an essential property of the clay it is made of.
- Unicity: which entities represent a whole and which entities are parts of these wholes? An *ocean* or *river* represents a whole but the *water* it contains does not.

The identity criteria are based on certain fundamental requirements. These include that the ontology is descriptive and reflects human cognition, perception, cultural imprints and social conventions (Masolo, Borgo,

Gangemi, Guarino, and Oltramari 2003).

The work of Guarino and Welty (2002a,b) has demonstrated that the WordNet hierarchy, when viewed as an ontology, can be improved and reduced. For example, roles such as AGENTS of processes are often non-rigid. They do not represent disjunct types in the ontology and complicate the hierarchy. As an example, consider the hyponyms of dog in WordNet, which include both types (races) like *poodle*, *Newfoundland*, and *German shepherd*, but also roles like *lapdog*, *watchdog* and *herding dog*. "Germanshepherdhood" is a rigid property, and a German shepherd will never be a Newfoundland or a poodle. But German shepherds may be herding dogs. The ontology would only list the *rigid* types of dogs (dog races):

Canine => PoodleDog; NewfoundlandDog; GermanShepherdDog, etc.

The lexicon of a language then may contain words that are simply names for these types and other words that do not represent new types but represent roles (and other conceptualizations of types). From this basic starting point, we can derive two types of mappings from synsets to the ontology (Fellbaum and Vossen 2007, Vossen and Fellbaum fc.):

1. Synsets represent disjunct types of concepts, where they are defined as:
   - names of Terms;
   - subclasses of Terms, in case the equivalent class is not provided by the ontology
2. Synsets represent non-rigid conceptualizations, which are defined through a KIF expression;

For example, English *poodle*, Dutch *poedel* and Japanse *pudoru* will become simple names for the ontology type: ⇔ ((instance x PoodleDog). On the other hand, English *watchdog*, the Dutch word *waakhond* and the Japanese word *banken* can be related through a KIF expression that does not involve new ontological types:

(and (instance, ?C, Canine),
      (instance, ?G Guarding)
      (role, ?C, ?G))

where we assume that *Guarding* will be defined as a process in the hierarchy as well in a future extension of MILO. The fact that the same expression can be used for all the three words indicates equivalence across the three languages.

The naming relation thus corresponds more or less with the way SUMO is currently mapped to the Princeton Wordnet, using equivalence and subsumption relations. The KIF expressions for non-rigid mappings are more similar to the axioms that found in SUMO, except that one of the variables in the axioms needs to correlate with the denotation of the synset that is being defined. In the case of the above example, the variable *?C* thus correlates with the possible referents of expressions with the syntactic head *watchdog*, *waakhond* and *banken*.

In a similar way, we can use the notions of Essence and Unicity to determine which concepts are justifiably included in the type hierarchy and which ones are dependent on such types. If a language has a word to denote a lump of clay (e.g. in Dutch *kleibrok* denotes an irregularly shaped chunk of clay), this word will not be represented by a type in the ontology because the concept it expresses does not satisfy the Essence criterion. Similarly, a Dutch word *rivierwater* (*river water*) is not represented by a type in the ontology as it does not satisfy Unicity; such words are dependent on other valid types through a more complex semantic relation.

5.2 Ontological implementation in Cornetto

The ontology mappings in Cornetto are currently restricted to triplets consisting of the relation name, a first argument and a second argument. It is thus not possible to represent complex KIF expressions as is done in the axioms of SUMO. However, by assuming default values for the KIF syntax, we can generate expressions that come close to these. The default operator of the triplets is AND, and we assume default existential quantification of any of the variables, specified as a value of the arguments. Furthermore, we follow the convention to use a zero symbol as the variable that corresponds to the denotation of the synset being defined and any other integer for other denotations. Finally, we use the symbol ⇔ for full equivalence (bidirectional subsumption). In the case of partial subsumption, we use the symbol ⇒, meaning that the KIF expression is more general than the meaning of the synset. If no symbol is specified, we assume an exhaustive definition by the KIF expression. The symbol ⇔ applies by default.

The following simplified expression can then be found in the Cornetto database for the above non-rigid synset of {waakhond} (watchdog):

(instance,0,Canine) (instance,1, Guarding) (role,0,1)

This should then be read as follows:

The expression exhaustively defines the synset (Û), AND there exists an instance 0 of the type Canine (instance, 0, Canine), AND any referent of an expression with the synset {waakhond} as the head is also an instance of the type Canine (the special status of the zero variable), AND there exists an instance of the type Guarding 1 (instance, 1, Guarding), AND the entity 0 has a role relation with the entity 1 (role, 0 ,1).

For names of types, we use the following expressions in Cornetto:

Hond (=, 0, Canine); the synset {hond} is a Dutch name

for the rigid type Canine

Bokser (+, 0, Canine); the synset {bokser} is a Dutch name for a rigid concept which is a subclass of the type Canine

Naming relations are mostly imported from the SUMO mappings to the English Wordnet through the equivalence relation of the Dutch synset to the English synset. In the case of {bokser}, the mapping is manually added because it is dog race that is not in the English Wordnet and not in SUMO. Possibly, SUMO could be extended with this Type.

Another case of mixed hyponyms are words for *water*. In the Dutch wordnet there are over 40 words that can be used to refer to water in specific circumstances or with specific attributes. Water is in SUMO a CompoundSubstance just as other molecules. We can thus expect that the synset of *water* in Dutch matches directly to Water in SUMO, just as *zand* matches to Sand. However, *water* has 3 major meanings in the Dutch wordnet: water as liquid, water as a chemical element and a water area, while there are only two concepts in SUMO: Water as the CompoundSubstance and a WaterArea. In SUMO there is no concept for water in its liquid form, even though this is the most common concept for most people. Most of the hyponyms of *water* in the Dutch Wordnet are linked to the liquid. To properly map them to the ontology, we thus first must map water as a liquid. This can be done by assigning the Attribute Liquid to the concept of Water as a CompoundSubstance. A SUMO axiom for this is:

```
(and (exists ?L ?W)
     (instance, ?W, Water) ,
     (instance, ?L Liquid)
     (attribute, ?L, ?W) )
```

In the Cornetto database, this complex KIF expression is represented by the simpler relation triplets:

(instance, 0, Water)(instance, 1, Liquid) (attribute, 1, 0)

The hyponyms of water in the Dutch Wordnet can further be divided into 3 groups:

- Water used for a purpose: theewater (*for making tea*), koffiewater (*for making coffee*), bluswater (*for extinguishing fire*), scheerwater (*for shaving*), afwaswater (*for cleaning dishes*), waswater (*for washing*), badwater (*for bading*), koelwater (*for cooling*), spoelwater (*for flushing*), drinkwater (*for drinking*)
- Water occurring somewhere or originating from: putwater (*in a well*), slootwater (*in a ditch*), welwater (*out of a spring*), leidingwater, gemeentepils, kraanwater (*out of the tap*), gootwater (*in the kitchen sink or gutter*), grachtwater (*in a canal*), kwelwater

(*coming from underneath a dike*), grondwater, grondwater (*in the ground*), buiswater (*on a ship*)
- Being the result of a process: pompwater (*being pumped away*), smeltwater, dooiwater (*melting snow and ice*), afvalwater (*waste water*), condens, condensatiewater, condenswater (*from condensation*), lekwater (*leaking water*), regenwater (*rain water*), spuiwater (*being drained for water maintenance*)

In Table 3, you find some of the mapping expressions that are used to relate these synsets to the ontology:

theewater (*tea water*)	**putwater** (*water in a well*)
(instance, 0, Water) (instance, 1, Tea) (instance, 2, Making) (hasPurpose, 1, 0) (resource, 0, 2) (result, 1, 2)	(instance, 0, Water) (instance, 1, MineOrWell) (located, 1, 0)
leidingwater (*tap water*)	**slootwater** (*in a ditch*)
(instance, 0, Water) (instance, 1, Device) (instance, 2, Removing) (origin, 0, 1) (patient, 0, 2)	(instance, 0, Water) (instance, 1, StaticWaterArea) (part, 0, 1)

Table 3: Triplets for some hyponyms of the Dutch water.

Through the complex mappings of non-rigid synsets to the ontology, the latter can remain compact and strict. Note that the distinction between Rigid and non-Rigid does not down-grade the relevance or value of the non-rigid concepts. To the contrary, the non-rigid concepts are often more common and relevant in many situations. In the Cornetto database, we want to make the distinction between the ontology and the lexicon clearer. This means that rigid properties are defined in the ontology and non-rigid properties in the lexicon. The value of their semantics is however equal and can formally be used by combining the ontology and the lexicon.

5.3 Ontology progress

The work on the ontology is mainly carried out manually. The mappings of the synsets to SUMO/MILO are primarily imported through the equivalence relation to the English wordnet. We used the SUMO-Wordnet mapping provided on: http://www.ontologyportal.org/, dated on April 2006. If there are more than one equivalence mappings with English wordnet, this may result in many to one mappings from SUMO to the synset. The mappings are manually revised traversing the Dutch wordnet hierarchy top-down so that we give priority to the most essential synsets. Furthermore, we will revise all synsets with a large number of equivalence relations or low-scoring equivalence relations. Finally, we also plan to clarify the synset-type relations for large sets of co-hyponyms as shown above for water. This work is still

in progress. We do not expect this to be completed for all the synsets in this 2-year project with limited funding but we hope that a discussion on this topic can be started by working out the specification for a number of synsets and concepts.

6. Conclusion

In this paper, we presented the Cornetto project that combines three different semantic resources in a single database. Such a database presents unique opportunities to study different perspectives of meaning on a large scale and to define the relations between the different ways of defining meaning in a more strict way. We discussed the methodology of automatic and manual aligning the resources and some of the differences in encoding word-concept relations that we came across. The work on Cornetto is still ongoing and will be completed in the summer of 2008. The database is freely available for research. The database and more information can be found on:

http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm

7. Acknowledgements

8. References

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

Fellbaum, C. and 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." Lecture Notes in Computer Science , Vol. 4568, eds. Ishida, Toru, Fussell, Susan R. and Vossen, Piek T. J. M. New York: Springer, pp. 1-16.

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

Guarino, N. and Welty, C., (2002). Identity and subsumption. In: R. Green, C. Bean and S. Myaeng (eds.), *The Semantics of Relationships: an Interdisciplinary Perspective*. Kluwer.

Guarino, N. and Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 45(2), 61-65.

Magnini, B., Cavaglià, G. (2000). Integrating subject field codes into WordNet. *Proceedings of the Second International Conference Language Resources and Evaluation Conference (LREC)*, Athens, Greece, 1413–1418.

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

Maks, I., Vossen, P., Segers, R., Vliet, H. van der, (2008). Adjectives in the Dutch semantic lexical database CORNETTO, this volume , LREC-2008, Marrakech.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari (2003). *WonderWeb Deliverable D18 Ontology Library*. Laboratory for Applied Ontology - IST-CNR, Trento, Italy.

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

Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada.

Niles, I. and Terry, A. (2004). The MILO: A general-purpose, mid-level ontology. *Proceedings of the International Conference on Information and Knowledge Engineering*. Las Vegas, Nevada.

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

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*, Szeged, Hungary, January 22-25, 2008.

Vossen. P., and Fellbaum, C. (to appear). Universals and idiosyncrasies in multilingual wordnets. In Boas, Hans (ed.), *Multilingual Lexical Resources*. Berlin: de Gruyter.

Vliet, H.D. van der (2007) The Referentie Bestand Nederlands as a multi-purpose lexical database. *International Journal of Lexicography* 20.3.