D03 Top-level ontology, relation constraints and assignments

Version: 1
Date: 07 May 2008

Project: STE05039
http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm

Funded by the Stevin framework (www.taalunieversum.org/stevin)

Authors:
- Piek Vossen, Irion Technologies/ Faculteit der Letteren, Vrije Universiteit Amsterdam
- Isa Maks, Faculteit der Letteren, Vrije Universiteit Amsterdam
- Roxane Segers, Faculteit der Letteren, Vrije Universiteit Amsterdam
- Hennie van der Vliet, Faculteit der Letteren, Vrije Universiteit Amsterdam

Table of Contents

Deleted: 18
Deleted: 27
Deleted: 28

1 Introduction

The Cornetto project (STE05039) is funded by the Nederlandse Taalunie in the STEVIN framework (www.taalunieversum.org/stevin). The goal is to build a lexical semantic database for Dutch, covering 40K entries, including the most generic and central part of the language. Cornetto will combine the structures of both the Princeton Wordnet and FrameNet for English, by combining and aligning two existing semantic resources for Dutch: the Dutch wordnet (Vossen 1998) and the Referentie Bestand Nederlands (Martin et al 1999). The Dutch wordnet (DWN) is originally based on the conceptual structure of the Van Dale VLIS database, and has been revised and refined in the EuroWordNet project (1996-1999, www.hum.uva.nl/~ewn). DWN has a similar structure as the English Wordnet although the top-level hierarchy was developed from an ontological framework and more horizontal relations are defined, such as roles and subevents. It has 55K entries, 70K word meanings and 110K semantic relations. The *Referentie Bestand Nederlands* (RBN) was developed with financial support of the governments of The Netherlands and Flanders (1996-1999, revised in 2004). It contains 45K entries with corpus-based descriptions, covering morphology, syntax, combinatorics, semantics and pragmatics. RBN is concept-based but also encodes the common word combinations that go with concepts, ranging from free combinations, to collocations, idioms, pragmatic formulae, proverbs and all sorts of typical and common usage and clichés, lexical functions and compositional and selectional preferences. Combinations can be encoded as lexical functions a la Me'lcuk or FrameNet-like structures.

The data for Cornetto will primarily be derived from conceptually aligning the two Dutch databases. Additionally, the database will be aligned to the Princeton Wordnet for English, from which further semantic resources can be accessed, such as ontologies and Wordnet Domains (Magnini et al. 2002). The alignment will be done automatically but will be refined in a manual editing process.

A new relation is the mapping from the synset to the ontology. 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* & a *dog*)

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

In this deliverable, we describe the work and decisions made for assigning a top-level ontology to the semantic structure of the Dutch database. The work was carried out as follows:

1. Study of literature and lexical-ontological models
2. Definition of a model for mapping a semantic language database to an ontology
3. Organisation of a workshop to discuss the ontology mappings
4. Extension of the database structure to incorporate the mappings
5. Import of the Wordnet-Sumo mappings to Cornetto
6. Revision of mappings during the editing
7. Extension of SUMO for adjective modelling

We defined the mapping relations between the ontology and the wordnet database. This is published and presented as part of the Global Wordnet Grid project (see publications and presentations). This model makes a clear distinction between the ontology and the synsets and allows for explicit mappings between synsets and the ontology through a naming relation, a subsumption relation or through complex expressions. These complex mapping relations are based on lists of triplets containing variables, SUMO terms and SUMO relations. The XSD of the database was extended to store the triplets.

We organized a project workshop on the 7th & 8th of June, were we invited the members of the user-group and the ontology-group for a discussion on the results so far. In addition to the user-group, we invited people and research groups that have expressed their interest in using the database. In total, about 30 people attended the workshop. We invited Adam Pease to give an invited talk about SUMO. On the 8th of June, we had a project meeting on the ontology assignment in Cornetto with Adam Pease. In this meeting we discussed our proposed model and we finalized our approach.

We created an improved mapping between the Dutch wordnet synsets and the English wordnet synsets. Through this more precise mapping, we could create an import that stores the equivalence mappings, the Wordnet domain labels and the SUMO mappings from the English Wordnet to the Dutch synsets. This import has been the starting point for the assignment of the SUMO labels to the Dutch wordnet.

The mappings have been revised for a number of critical cases. Furthermore, we worked out a number of complex mappings. These were also presented and discussed on the workshop.

The adjectives are difficult to account for in Cornetto. In English Wordnet, there is no systematic classification of adjectives. In the Dutch wordnet they have only marginally been encoded and SUMO mostly classifies adjectives in Wordnet as SubjectiveAssessmentAttribute. We therefore made an extension of SUMO on the basis of a linguistic typology of adjectives.

.

This document is further organized as follows. In the next section, we first describe the rationale behind the ontologization of the Cornetto database. In section 3, we motivate our choice for the ontology SUMO/MILO and in section 4, we explain some ontological assumption that we have adopted. Section 5 describes how we implemented these in the Cornetto database. Nex in section 6, we describe the work on extending the ontology for adjective classification and section 7 reports on the first results of consistency checking through the ontology.

The work for this WP has partially been published in various reviewed papers:

A. Pease, C. Fellbaum, P. Vossen, (fc)
"Building the Global WordNet Grid", in: Proceedings of the 18th International Congress of Linguists (CIL18), Seoul, Republic of Korea, July 21-26, 2008.

Vossen P., Fellbaum C. (fc) "Universals and Idiosyncracies in Multilingual WordNets", in: Handbook Multilingual Lexicography, Oxford University Press, 2008

Vossen P., Maks I., Segers R., VanderVliet H. (fc) "Integrating lexical units, synsets and ontology in the Cornetto Database", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30 May 2008.

Maks I., P. Vossen, Segers R., VanderVliet H., van Zutphen H. (fc) "Encoding adjectives in the Dutch semantic lexical database Cornetto", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30 May 2008.

Fellbaum C. and P. Vossen (2007), "Connecting the Universal to the Specific: Towards the Global Grid", In: Proceedings of The First International Workshop on Intercultural Collaboration (IWIC 2007), Kyoto, Japan, January 25-26, 2007 Publically available on external Website

Vossen, P. (2006), *Een communicatief lexicon door het verankeren van woordbetekenissen*, Oratie, leerstoel Computationele Lexicologie, Vrije Universititeit Amsterdam. (PDF)

2 Rationale for a ontological layer in the Cornetto database

The ontology in Cornetto represents a language neutral anchoring of the meaning of Dutch words in a logical formal framework. This idea was further explored in Vossen 2006. Such a framework has 3 important advantages:

1. It represents a more formal and fundamental model of meaning that can be used as a basis for making decisions about the different meanings encoded in RBN and DWN;
2. It makes it possible to detect inconsistencies in the Dutch database, i.e. the semantic relations between Dutch word meaning need to be consistent with the constraints in the ontology;
3. It makes it possible to use the semantic database in a formal system that needs to make inferences, i.e. computer programs can reason on Dutch text in the same way as for other languages;
4. It makes it possible to more easily relate wordnet databases and other semantic lexicons across languages, thus establishing a global semantic grid of meaning across languages and cultures;

The first motivation is discussed in the various papers (Horak et al 2008fc, Vossen et al 2008a,b) on the Cornetto database architecture and in deliverable *D01 Design of the Cornetto database*. RBN and DWN have different starting points for the distinguishing meanings. For merging the distinctions in a unified model, it is important to have explicit and independent criteria. The ontology and ontological principles can be a good instrument for this. Especially, in the case of merging adjectives, this has been important due to the lack of a common framework. This is further discussed in section 5.

Atserias et al. 2008 demonstrated how wordnets can be improved and checked using an ontological framework. They used the EuroWordNet top-level ontology that was assigned to the so-called Common Base Concepts (Vossen 1998). These Base Concepts are important in a series of wordnets. The top-level ontology consists of 63 semantic types, where each type expresses a basic semantic class

of concepts, e.g. Animate versus Artifact, Part versus Group versus Whole, etc. These distinctions are taken from other semantic frameworks.

Through the hyponymy relations in wordnet, the ontology is inherited to more specific concepts in a wordnet. This has revealed 227,908 incompatibilities, i.e. combinations of ontology types that are incompatible, e.g. synsets labelled both as Object and Substance. Atserias et all (2008) could repair these by 13,613 changes (adding or removing ontology assignments) in the ontology assignment and 359 so-called blockage points. Blockage points are used in cases where the wordnet hierarchy is inconsistent but could not be changed. In that case, the inheritance is blocked and another top-level concept is assigned to the lower levels. A similar approach can be applied to the Cornetto database. It can be used to guide the editing and verify the quality of sections of the database. This is further discussed in section 6.

The second point is related to semantic web applications that need to handle textual data in a formal environment or information system. In this case, the semantics of words in a language needs to be mapped to a formal model. We foresee these applications in the near future, especially when integrated with domain modelling of ontologies reusing general knowledge and vocabularies. A similar research was demonstrated by Tufis et al (2006), who apply word-sense-disambiguation of Romanian text through SUMO linked to the Romanian wordnet.

From a cross-linguistic perspective, the anchoring of semantic lexicons in languages opens many possibilities for contrastive studies, for shared formal reasoning, for cross-lingual applications, etc. These ideas have cumulated in the idea of the Global Wordnet Grid (Fellbaum and Vossen 2007, Vossen and Fellbaum fc., Pease, Fellbaum and Vossen 2008fc). The Global Wordnet Grid project foresees that wordnets for all languages in the world are not interconnected through English but through a formalized ontology. Various wordnets have already been mapped to SUMO and MILO: English wordnet (Niles and Pease 2003, Niles and Terrry 2004), the Arabic Wordnet (Fellbaum et al 2006, Elkateb et al, 2008), the Romanian wordnet (Tufis et al 2006) and the Chinese wordnet (Huang et al fc.). Linking Cornetto to the same ontology establishes a further step in the direction of the GWG.

Finally, the KYOTO project (funded as project 211423 in the 7th Frame Work in the area of Digital Libraries: FP7-ICT-2007-1, Objective ICT-2007.4.2: Intelligent Content and Semantics) can be seen as a further implementation of these ideas (Vossen et al 2008c, fc). In the KYOTO project, we will further explore and re-use the ontological mappings and relations from Cornetto and share this with other wordnet builders in the project.

3 Selection of ontological framework

In the Cornetto project, we have selected SUMO/MILO as the ontological framework. The motivation for this choice is mainly pragmatic:

1. The main purpose of the ontology is to help and support the development of the Dutch lexical semantic database;
2. There is no large-scale ontology without dispute;

3. The selected ontology should have a sufficient-coverage to cover the wide range of meanings and distinctions in a large-scale lexicon;
4. It should be possible to import the ontology on a large scale to the Cornetto database;
5. The ontology should establish strong relations with the NLP community;

Mascardi, Cordì, and Rosso (2008) review 7 different Upper Ontologies: BFO, Cyc, DOLCE, GFO, PROTON, Sowa's ontology, and SUMO, according to a set of standard software engineering criteria. Their evaluation involves the following criteria:

- Number of Dimensions.
- Implementation language(s) A fragment of OWL Lite.
- Modularity.
- Use in Applications.
- Alignment with WordNet.
- Licensing.

Their comparison is objective and descriptive. There is no ranking of the ontologies. The results are summarized in Table 1, based on their technical paper. More details can be found in the original paper.

Deleted: Table 1

The most important features for Cornetto are coverage and linking to the English wordnet. Formal ontological considerations and disputes are less relevant. First of all, we are not directly using the ontology for complex inferencing but merely for checking the semantic implications of the Cornetto database as such.

Furthermore, we believe that deciding on these issues is out of the scope of this project and should be solved in the ontology community. An accepted solution in that community can then be applied to the Cornetto project at a later point and could possibly be helped through a large coverage ontology that is linked during the project. We also believe that fundamental changes will apply to the high and core levels of the database and that very specific labels are rather uncontroversial.

Given the availability, size, coverage and the linking to Wordnet, the obvious candidate for our purpose is SUMO/MILO. Other ontologies have a too narrow scope and coverage or are not linked to Wordnet and freely available. Linking to SUMO/MILO also provides us with a good mapping to other wordnet-like projects such as Arabic Wordnet, Romanian Wordnet, Chinese Wordnet and the Global Wordnet Grid.

Nevertheless, we decided to adopt the ontologization principles from the OntoClean method (Guarino and Welty (2002a, 2002b). In the case that semantic implications are not straight forward, these principles will be followed for revisions. A similar approach was followed in the SmartWeb project. Oberle et al (2007) compared Dolce versus Sumo from the perspective of a domain ontology in their SmartWeb application. Again, the application is here the major motivation for the evaluation. Their main conclusion is that the design principles of Dolce are useful for developing a domain specific ontology for an application, but Sumo has a larger coverage and hierarchy. Their combination (SmartSUMO) likewise combines the best of both worlds.

Table 1: Ontology comparison table based on Mascardi, Cordì, and Rosso

	Home page	Developers	Dimensions	Modularity	Applications	Alignment wordnet	Language	License
BFO	http://www.ifomis.org/ bfo	Smith, Grenon, Stenzhorn, Spear (IFOMIS)	36 classes related via is_a relation,	Two modules: SNAP: snapshot ontologies indexed by times & SPAN: single videoscopic ontology	Applied to the biomedical domain and used in building an ontology for clinic-genomic trials on cancer.	NO	OWL	Free
Cyc	http://www.cyc.com/	Cycorp	300,000 concepts, 3,000,000 assertions (facts and rules), 15,000 relations	The "microtheory" approach supports modularity	Domains of NLP, e.g.: WSD and Q&A, network risk assessment, terrorism-related	Links to 12,000 synsets	CycL, OWL	Commercial, OpenCyce for research
DOLCE	http://www. loa-cnr.it/ DOLCE.html	Guarino et al. of the LOA	100 concepts 100 axioms	It is not currently divided into modules (planned).	LOIS Project, SmartWeb, Language Technology for eLearning AsIsKnown	Links to 100 synsets	First Order Logic, KIF, OWL	Free
GFO	http://www.onto-med.de/ontologies/gfo.html	Onto-Med Research Group	79 classes, 97 subclassrelations, 67 properties	3-layered architecture: abstract top level, abstract core level, and basic level. Several ontological modules, incl.	Ontological foundation of conceptual modelling and Biomedical science: Gene Ontology, Celltype Ontology, Chemical Entities of Biologic al Interest Ontology, GFO-Bio.	NO	First Order Logic and KIF (forthcoming); OWL	released under the modified BSD Licence

				functions and roles				
PROTON	http://proton.semanticweb.org/	Ontotext Lab, Sirm	300 concepts 100 properties	3 levels including 4 modules.	Different domains and purposes, e.g. semantic annotation, knowledge management systems in legal and telecommunications domain (projects MediaCampaign, ISTWorld, Business Data Ontology for Semantic Web Services)	NO	OWL Lite	Free
Sowa's	http://www.jfsowa.com/ontology/	Sowa	30 classes, 5 relationships, 30 axioms	Not explicitly divided into modules	Inspired many other upper ontologies,	NO	1st Order Modal Language,KIF	Free
SUMO	http://www.ontologyportal.org/	Niles, Pease, Menzel	20,000 terms 60,000 axioms (incl.domain ontologies)	MId-Level Ontology, and ontologies for a range of specialized domains	Many papers report on usage (from academic to government, to industrial), among which NLP, "pure" representation and reasoning.	All synsets of WN3.0	SUO-KIF, OWL	Free

4 Ontological assumptions

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 and Welty (2002a, 2002b), identity criteria can be used to determine the set of disjunct 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 anti-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, 2002b) 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 anti-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.

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 Implementation in Cornetto

The ontology mappings in the Cornetto database are 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), just as with the SUMO wordnet mappings. 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 partial subsumption.

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 (=), 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 a dog race that is not in the English Wordnet and not in SUMO. Possibly, SUMO could be extended with this Type.

When we look at the different dogs in the Dutch wordnet then we see 3 types of hyponyms:

- bokser; corgi; loboor; mopshond; pekinees; pointer; spaniël;
- pup; reu; teef
- bastaard; straathond; blindengeleidehond; bullebijter; diensthond; gashond; jachthond (hunting dog); lawinehond; schoothondje (lap dog);waakhond (watch dog)

The first group are names for doc races that are clearly rigid and disjunct. They represent names for Terms. The second group words for male/female and baby dogs. They can be encoded in the same way as man, woman and child for humans. The third group refers to dogs with certain non-rigid attributes. They will thus not represent names for types but are related to the ontology by a mapping to the term Canine and the attribute that applies.

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

Deleted: Table 2

Table 2: Triplets for some hyponyms of the Dutch water

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)

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.

6 Revising SUMO for encoding adjectives

Automatic alignment performed relatively poorly in case of adjectives, since each resource has its own system of defining word senses, focusing on different distinctions.

In SUMO, all attribute classes are subsumed by either RelationalAtribute or Internal Attribute. Relational Attribute is defined as “Any attribute that an entity has by virtue of a relationship that it bears to another Entity or set of Entities”. Typical examples are those attributes which are somehow related to social or civil systems, like for instance

- ReligiousAttribute (e.g. *catholic, religious*)
- SocialRole (e.g. *teacher, noble*)
- SocialPosition Attributes (e.g. *rich*)
- Etc.

The InternalAttribute Class is defined as “Any Attribute of an Entity that is an internal property of the Entity, e.g. its shape, its colour, its fragility, etc.”. Typical subclasses are:

- ShapeAttribute (e.g. *round*),
- SizeAttribute (*wide*),
- LengthAttribute (e.g. *short*),
- BiologicalAttribute (*hungry*)
- PhysicalAttribute (e.g. *dense, wet*),
- PerceptualAttribute (e.g. *loud, aromatic, sour, rough*) and
- EmotionalState (e.g. *happy, angry*)

- Etc.

In order to check if the existing SUMO would be complete and fine-grained enough to account for the semantics of general language adjectives, we compared it with a richer adjective classification. From the various classification systems that have been developed (e.g. like the Mikrokosmos approach, Raskin et al. 1995, the SIMPLE approach, Peters et al. 2000), we chose the semantic classification developed for German and applied in the German Wordnet (Hundschnurscher and Splett 1982, H&S from now on). This semantic classification is based on a traditional lexicographical approach to word sense disambiguation, which appeared to fit closely to the approach underlying DWN and RBN. In this classification, adjectives are divided into seventy semantic classes, each referring to common characteristics of sets of nouns. The classes are organized in fifteen main classes which roughly coincide with the semantic typology used in RBN (see 2.1).

To test the usability and effectiveness of the classification, we tried to map the 100 most-frequent adjectives (which we already manually mapped on the H&S classification) with the SUMO. When comparing the H&S classification with SUMO, many main classes could be easily transferred (see the following Table)[1].

Table 3: Comparison of H&S semantic classes and SUMO terms

H&S	SUMO
Perception-related	PerceptionalA
Material-related	PhysicalA
Body-related	BiologicalA
Mood-related	EmotionalA
Character-Behaviour	TraitA
Spirit-related	-
relationalA	RelationalA
General /evaluative	NormativeA
Temporality-related	-
Weather-related	Temperature
Social-Related	RelationalA

However, with regard to subclasses we encountered several problems due to gaps and inconsistencies in SUMO and due to inconsistencies between the two systems. Most problems were solved by changing and expanding the SUMO ontology (see figure 1). We present some of the problems in more detail:

- Gaps

Some important H&S main and subclasses are missing from the ontology. Among them are adjectives which refer to time (*short*), adjectives which refer to cognitive features (*intelligent, acute, shrewd*) etc. We added these to SUMO.

[1] We cannot always be sure if these classes really overlap since the H&S classification defines them by giving typical examples only and SUMO defines them by giving a formal definition.

- Unbalancedness

On various levels, SUMO is unbalanced with respect to the fine-grainedness of the classes. For example ConsciousnessAttribute *(conscious, semi-conscious)* is directly subsumed by BiologicalAttribute and therefore on the same level as a higher level class like Psychological Attribute. In this particular case we introduced the class BodilyAttribute which subsumes all kinds of attributes referring to the human or animal body.

Other examples of this kind are classes which refer to rather specific material-related attributes like Saturation Attribute (e.g. wet, dry) and BreakabilityAttribute (e.g. fragile, robust). Both are directly subsumed by InternalAttribute and therefore on the same level as PhysicalAttribute. We think they should be lower in the hierarchy and, therefore, we moved them.

- Inconsistencies between the two systems

We encountered inconsistencies with regard to classes of comparable concepts which were integrated in the ontology at completely different levels. For instance, in SUMO the class of PerceptualAttributes has the following four subclasses:

OlfactoryAttribute (*e.g. aromatic*)
TasteAttribute (e.g. *sweet)*,
VisualAttribute (e.g. *transparent*)
TactileAttribute (e.g. *smooth*).

Typical definitions of these subclasses are: “the class of attributes relating to the taste of objects” for TasteAttribute , “the of class of properties that are detectable by smell’’ for OlfactoryAttribute. However, attributes that relate to the sound of objects, i.e. the class of SoundAttribute, are in SUMO not considered as an InternalAttribute but as a RelationalAttribute and defined as “the volume of sound relative to a listener”.

In H&S the SoundAttribute is classified together with other perceptual attributes regarding color, smell, touch, etc. Deciding how to resolve inconsistencies between the two systems was not always straightforward. The SUMO rationale for distinguishing SoundAttribute from the other perceptual attributes is that sound is considered a RelationalAttribute: a string, for instance, does not produce sound, unless it is brought into motion by something else. In this sense, the sound-producing property of the string may be considered as not strictly intrinsic and is therefore not considered as an InternalAttribute. However, the same holds for color. An object may have the property of being red, but this property becomes manifest only when light reaches the object. In the absence of light, there is no color. Thus, a property that is seemingly truly ‘internal’ such as color, may be considered relational as well.

The example shows that contrasting the two systems requires resolution of certain questions that are not completely unambiguous. Although in a practical sense, this poses a problem, conceptually it is interesting, because it forces us to reflect on the nature of the adjectives. Hopefully, this will help to develop an ontology that allows for productive computerized reasoning. At this stage of our project, it is difficult to predict which resolutions of conflicts between SUMO and

H&S should be preferred. In this particular case, we chose to follow H&S and to keep together the five perceptual attributes.

The result of the comparing and merging of SUMO and the H&S classification is shown in a corrected and extended attribute branch of SUMO (figure 1). The hierarchy consists of:

- Original SUMO concepts as far as needed for the description of the 100 most frequent adjectives.
- Moved SUMO concepts (marked with +)
- New concepts from the H&S classification (marked with ++).

The structure represents the semantics for the 100 most-frequent adjectives. It is still a preliminary classification; as we proceed we expect to add more classes and subclasses to cover domain-specific attributes.

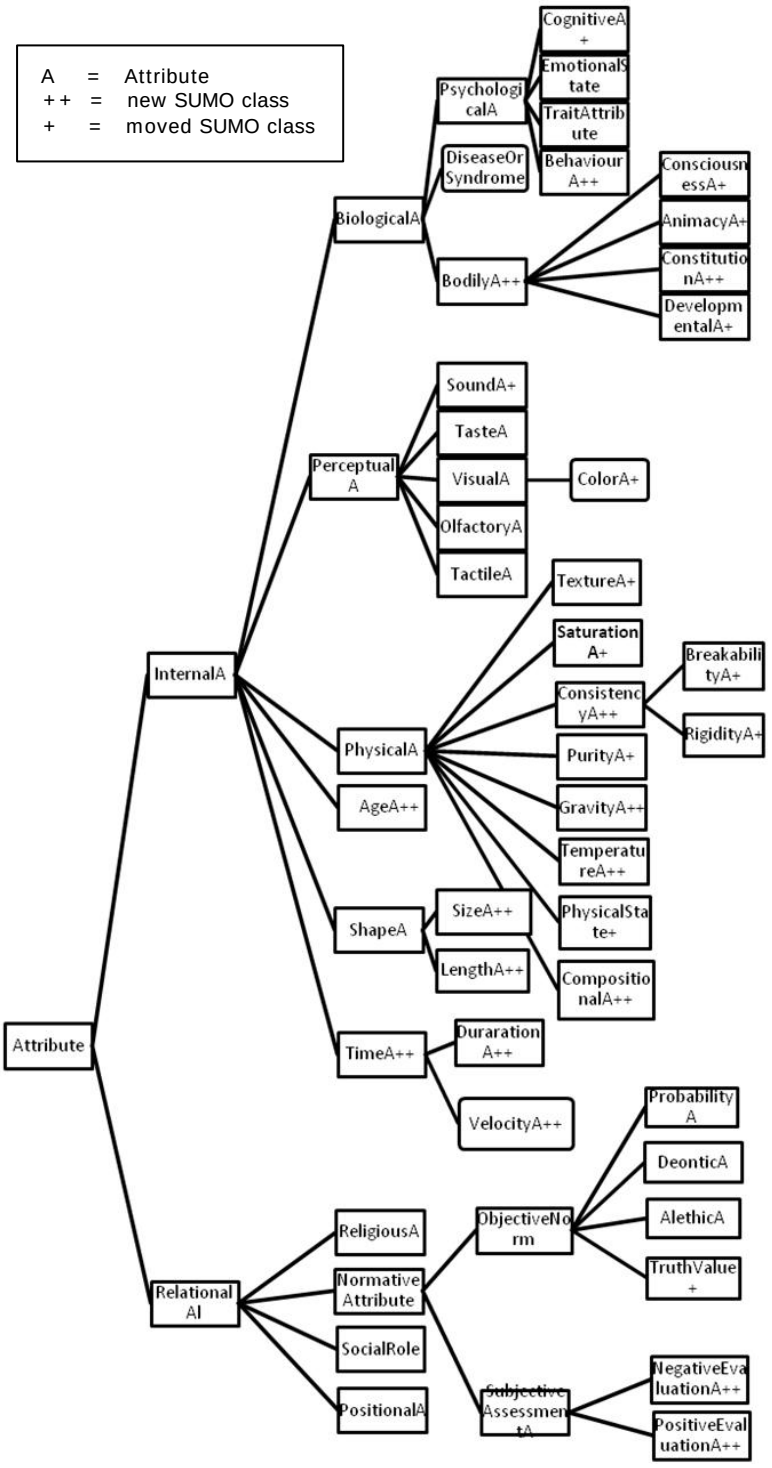

Figure 1 SUMO for CORNETTO adjective concepts

7 Mapping synsets to the ontology

The mapping synsets to the ontology is mainly carried out manually, when Lexical Units are also aligned to synsets. This work is concentrated on the most-polysemous and most-frequent words. The mappings of the synsets to SUMO/MILO are primarily imported through the equivalence relation to the English wordnet. Also the mapping relation is copied from the SUMO to English Wordnet relation.

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. 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.

Nevertheless, we will apply consistency checking of ontology assignments to the database. Whereas Atserias et al (2008) measure the direct mapping of the EuroWordnet Top-Ontology, we can only indirectly evaluate the mapping based on the mapping to the English Wordnet. Obviously, the import will be better after the mappings to English have been revised. Without such a revision, we can also consider the synsets with many ontology mappings, as is shown in Table 4.

Deleted: Table 4

Table 4: Sumo/Milo mappings imported through English Wordnet relations

NR OF SUMO RELATIONS	zero	one	two	three	four or more
NOUNS	42240	54829	6356	2191	1455
VERBS	806	4799	2096	822	598
ADJECTIVES	2293	3123	1392	515	667

From this table, we can conclude that more than 60% of the nouns and adjectives and even more verbs have been mapped to SUMO through the English Wordnet. Note that not all words have an equivalence mapping to English. Only a small proportion has a large amount of Sumo Mappings. Below in Tables 5 to 7 are lists of the Sumo Type combinations that are projected on the Dutch synsegts.

As a first evaluation of consistency, we extracted the compatibility of multiple SUMO mappings. The consistency was defined by the average hierarchical distance of each mapping to all other mappings, where distance is measured by the number of steps up in the SUMO hierarchy to find a matching Type in the parent tree of each mapping.

(i) $S(m_{i,j}) = (n \mid m_i = m_j)$

(ii) $D(m_{i,j}) = Sm_{ij} + Sm_{j,j}$

(iii) $\frac{SUM(D(m1..n)}{N}$

According to (i), S represents the number of steps for a SUMO Type to match another SUMO type by traversing the hierarchy. The distance D is the number of steps up the hierarchy of one Type plus the number of steps up the hierarchy of another Type. The third formula is the average value for D given all the Types assigned to a synset.

The tables Table 5, Table 6 and Table 7 illustrate the effect of the tree distance score for nouns, verbs and adjectives with 4 or more SUMO mappings. For each POS, we gave a few examples with minimum and maximum distance.

Deleted: Table 5

Deleted: Table 6

Deleted: Table 7

The first example in the noun table is *gedierte* (unwanted animals), which has the SUMO classes: Animal:0.5, Arthropod:0.75, Invertebrate:0.25, Vertebrate:0.5. Behind the SUMO class, a score indicates the average distance of the class to all the other classes. Interesting enough, it has incompatible classes that are still very close. Actually, the word can be used to refer to any type of animal that is causing trouble. The second example is *vuurbuis* (end of a gun), which has four classes that are all incompatible. Some of them are redundant, being the parent of other classes. The third example *bank* (sand bank) has the classes: GeographicArea:0.5, LandArea:0.25, ShoreArea:0.75, WaterArea:0.5. These all make sense since the word stands for banks of sand or rocks at the shores of the coast. The average score for all these cases is 0.5 and the classes are indeed mostly compatible. Incompatibilities as between Vertebrate and Invertebrate can be filtered out separately by using the DisjunctSet classes in SUMO.

At the end of the table, we see examples with the largest average distance, thus most incompatible. We not only see that the distance is over 9 on average but also that they tend to have more than 4 mappings. The last example *rederij* (ship carrier company) is a company (Corporation), which is logically incompatible with the 3 other classes: AircraftCarrier, Position and Possesses. The example *godsdiensplechtigheid* (religious service) has many classes that are extremely incompatible.

For verbs and adjectives, we see a similar difference between classes with low distance and high distance.

Table 5: Nouns with 4 or more SUMO mappings scored by tree distance

Word meaning	SUMO Count	Scored Pattern	Average Score
gedierte:2	4	Animal:0.5.Arthropod:0.75.Invertebrate:0.25.Vertebrate:0.5.	0,5
vuurbuis:1	4	Artifact:0.5.Device:0.25.EngineeringComponent:0.75.Tube:0.5.	0,5
bank:3	4	GeographicArea:0.5.LandArea:0.25.ShoreArea:0.75.WaterArea:0.5.	0,5
bewijs:2	4	Certificate:0.25.FinancialInstrument:0.5.Text:0.75.Ticket:0.5.	0,5
stootkussen:2	4	Artifact:0.5.Device:0.25.EngineeringComponent:0.75.StationaryArtifact:0.5.	0,5
markies:3	4	Artifact:0.5.Device:0.5.Porch:0.75.StationaryArtifact:0.25.	0,5
dekkleed:1	4	Blanket:0.5.CottonFabric:0.5.Device:0.75.Fabric:0.25.	0,5
transmissie:4	4	Device:0.5.DriveComponent:0.75.EngineeringComponent:0.25.Machine:0.5.	0,5
wagenboom:1	4	Artifact:0.5.Device:0.25.EngineeringComponent:0.75.StationaryArtifact:0.5.	0,5
stoof:3	4	Device:0.5.Fireplace:0.5.Machine:0.5.Stove:1.	0,62
deurvleugel:1	4	Artifact:0.75.Building:0.75.Door:0.75.Room:0.75.	0,75
vuurwater:1	4	AlcoholicBeverage:0.5.Beverage:0.5.DistilledAlcoholicBeverage:1.Food:1.	0,75
alcoholica:1	4	AlcoholicBeverage:0.5.Beverage:0.5.DistilledAlcoholicBeverage:1.Food:1.	0,75
logboek:2	4	Artifact:1.Book:0.75.Periodical:1.25.Text:0.5.	0,88
roosterwerk:1	4	ArtWork:0.75.Artifact:0.5.Device:0.75.SelfConnectedObject:1.5.	0,88
slurf:1	4	BodyPart:0.5.Organ:0.75.PlantBranch:0.75.Torso:1.5.	0,88
strooier:1	4	Artifact:0.75.Clothing:1.25.Device:0.5.TransportationDevice:1.	0,88
stuk:10	4	ArtilleryGun:0.75.Firearm:0.75.Gun:0.5.Weapon:1.5.	0,88
stellage:3	4	Artifact:0.75.Clothing:1.25.PerformanceStage:1.StationaryArtifact:0.5.	0,88
indispositie:1	4	DiseaseOrSyndrome:1.EmotionalState:1.25.PsychologicalAttribute:0.5.TraitAttribute:0.75.	0,88
kelder:1	4	Artifact:1.Basement:1.25.Room:0.75.StationaryArtifact:0.5.	0,88
verificatie:1	4	Investigating:0.75.Measuring:1.25.Pursuing:1.25.Reasoning:0.75.	1
zelfvoldoening:1	4	EmotionalState:0.75.Happiness:1.25.PsychologicalAttribute:1.TraitAttribute:1.	1
zitbank:1	4	Artifact:1.Device:1.Seat:0.75.Sofa:1.25.	1
closet:1	4	Building:1.Device:0.75.Room:1.Toilet:1.25.	1
schuldbrief:1	12	Artifact:7.Bond:9.92.Certificate:8.25.Contract:10.67.Device:7.67.Obligation:9.Radiating:7.92.Relation:5.92.TextureAttribute:8.08.connects:8.25.holdsObligation:9.42.holdsRight:9.42.	8,46

blaag:1	15	AnatomicalStructure:7.33.Anxiety:12.13.Ape:7.47.Artifact:6.6.CognitiveAgent:6.6.Fish:8.33.Foot:9.07.Human:7.47.Male:9.53.Monkey:7.47.Position:8.73.Primate:9.47.Shoe:9.2.SocialRole:7.87.believes:10.2.	8,5
not_found:0	4	Law:8.75.Share:7.25.Sharing:8.75.financialAsset:9.25.	8,5
tegemoetkoming:2	6	Committing:8.67.Contract:10.67.CurrencyMeasure:8.Giving:7.17.UnilateralGiving:7.83.subProcess:8.67.	8,5
tweede:3	11	BodyJunction:11.64.DramaticActing:10.PolyphonicMusic:10.82.Region:7.55.TimeInterval:10.45.agent:6.91.financialAsset:7.91.holdsObligation:7.09.part:7.45.properPart:7.09.subProcess:6.91.	8,53
opsteker:1	8	Buying:7.62.Contract:11.5.FruitOrVegetable:10.5.IntentionalProcess:5.62.Requesting:9.75.Selling:7.62.SubjectiveAssessmentAttribute:8.5.UnilateralGiving:7.38.	8,56
kattenkop:2	6	Device:6.83.Expressing:8.Feline:10.67.Female:9.LandVehicle:8.83.SubjectiveAssessmentAttribute:8.33.	8,61
vrijstelling:3	6	Certificate:9.33.Contract:9.83.NormativeAttribute:6.5.RegulatoryProcess:7.67.Removing:8.33.holdsRight:10.	8,61
kippenkontje:1	7	BackFn:7.86.BodyPart:9.71.EngineeringComponent:9.71.Position:8.29.TimeInterval:9.71.part:7.86.superficialPart:7.43.	8,65
los:4	19	Article:8.53.Artifact:6.84.BaseballSingle:9.74.Clothing:8.53.ContentBearingObject:6.47.Device:7.63.Eating:10.63.Feline:12.Human:8.42.Illuminated:9.11.Position:8.37.PositiveInteger:11.47.RadiatingLight:7.68.RadiatingVisibleLight:8.58.Region:5.84.SubjectiveAssessmentAttribute:8.37.VisualAttribute:8.21.believes:9.26.knows:9.26.	8,68
contingent:2	6	Giving:8.17.Law:10.17.Share:8.17.Sharing:8.17.financialAsset:9.17.part:8.5.	8,72
uitzuiging:1	4	FinancialTransaction:8.75.Ordering:8.75.SubjectiveAssessmentAttribute:7.75.resource:9.75.	8,75
volmacht:1	5	Certificate:9.PoliticalOrganization:9.SubjectiveAssessmentAttribute:7.8.TimeInterval:9.holdsRight:9.6.	8,88
afbetaling:1	8	Certificate:9.88.Contract:11.88.Horsepower:8.88.Payment:8.38.Removing:8.38.Selecting:7.12.Selling:8.38.price:9.62.	9,06
factuur:1	4	Advertising:8.5.CurrencyMeasure:8.FinancialBill:10.5.Law:9.5.	9,12
verkoop:1	4	Contract:10.CurrencyMeasure:8.Offering:9.Selling:9.5.	9,12
godsdienstplechtigheid:1	16	Cooperation:6.75.EmotionalState:11.06.Giving:7.5.Impacting:9.31.IntentionalProcess:5.5.MilitaryService:11.19.ReligiousProcess:7.ReligiousService:10.19.Repairing:6.12.ServiceOrganization:9.44.Serving:10.12.SexualReproduction:10.19.SubjectiveAssessmentAttribute:9.31.Tableware:9.94.employs:9.94.instrument:12.56.	9,13
rederij:1	4	AircraftCarrier:11.25.Corporation:7.75.Position:8.25.possesses:9.75.	9,25

Table 6: Verbs with 4 or more SUMO mappings scored by tree distance

Word meaning	sumo Count	Scored Pattern	Average Score
kuipen:1	4	ContentDevelopment:0.5.IntentionalProcess:0.5.IntentionalPsychologicalProcess:0.25.Planning:0.75.	0,5
vreten:4	4	IntentionalPsychologicalProcess:0.5.Keeping:0.5.Pursuing:0.5.RegulatoryProcess:1.	0,62
omrollen:4	5	DirectionChange:0.6.Motion:0.6.Reversing:0.6.Rotating:0.6.Transfer:1.2.	0,72
traceren:1	4	IntentionalProcess:0.75.Investigating:0.75.Planning:0.75.Pursuing:0.75.	0,75
omzetten:5	6	DirectionChange:0.83.Motion:0.83.Returning:1.Rotating:0.83.Transfer:1.Translocation:0.5.	0,83
renderen:1	4	Buying:0.75.FinancialTransaction:0.5.Giving:1.5.Payment:0.75.	0,88
omwroeten:1	4	IntentionalProcess:0.5.IntentionalPsychologicalProcess:1.Motion:0.75.ShapeChange:1.25.	0,88
hergroeperen:1	4	Combining:1.25.IntentionalProcess:0.5.Motion:0.75.OrganizationalProcess:1.	0,88
uitbetalen:2	4	Buying:0.75.FinancialTransaction:0.5.Giving:1.5.Payment:0.75.	0,88
bijstorten:2	4	Buying:0.75.FinancialTransaction:0.5.Giving:1.5.Payment:0.75.	0,88
vastklemmen:3	4	Attaching:5.5.Ordering:7.5.Touching:6.5.grasps:9.	7,12
rondsjouwen:2	4	Contest:6.Requesting:7.5.Transportation:6.attribute:9.	7,12
blazen:2	5	ExpressingDisapproval:7.2.OrganismProcess:6.8.SoundAttribute:7.2.Transfer:6.2.Wind:9.	7,28
aankunnen:3	7	Managing:8.57.Ordering:10.SubjectiveAssessmentAttribute:7.71.capability:5.86.hasPurpose:6.71.needs:5.86.part:6.71.	7,35
aanhouden:9	30	BiologicalAttribute:7.47.Buying:9.03.Decreasing:6.63.Eating:9.43.Getting:7.17.Imprisoning:7.3.Increasing:6.63.IntentionalProcess:4.77.IntentionalPsychologicalProcess:5.47.Keeping:5.43.Maintaining:5.47.MotionUpward:5.77.Obligation:10.13.PastFn:8.Perception:7.57.Process:4.63.Putting:7.43.SocialInteraction:5.4.Speaking:8.93.SubjectiveAssessmentAttribute:8.27.Touching:7.43.Transportation:6.57.WhenFn:8.Writing:6.4.believes:7.93.duration:9.53.employs:7.93.holdsRight:9.53.possesses:9.53.prevents:7.93.	7,39
schuilgaan:2	6	Giving:7.33.LyingDown:6.67.StateChange:6.TellingALie:8.67.causes:8.33.located:7.67.	7,44
inblazen:1	5	Obligation:9.Requesting:7.8.Stating:7.2.causes:8.4.increasesLikelihood:7.2.	7,92

Table 7: Adjectives with 4 or more SUMO mappings scored by tree distance

Word meaning	sumo Count	Scored Pattern	Average Score
edel:1	4	NormativeAttribute:0.SocialRole:0.25.SubjectiveAssessmentAttribute:0.25.eh-b/c(+:-0.5.	
vrolijk:1	4	PsychologicalAttribute:0.StateOfMind:0.TraitAttribute:0.	
stuurvast:1	4	PsychologicalAttribute:2.SubjectiveAssessmentAttribute:1.5.True:1.5.forall:-0.5.	
gebroken:6	4	Function:1.75.SubjectiveAssessmentAttribute:3.25.agent:2.forall:-0.5.	
zedenbedervend:1	4	NormativeAttribute:1.SocialRole:1.25.SubjectiveAssessmentAttribute:1.5.TextureAttribute:2.75.	
constant:1	4	Quantity:1.75.SubjectiveAssessmentAttribute:2.5.capability:2.75.forall:-0.5.	
suizelig:1	6	ConsciousnessAttribute:2.17.DiseaseOrSyndrome:1.5.PsychologicalDysfunction:2.StateOfMind:1.67.SubjectiveAssessmentAttribute:3.67.body-related:affliction(-negative:-0.33.	
expressief:2	4	SubjectiveAssessmentAttribute:3.5.and:-0.5.capability:2.25.instance:2.25.	
ingebonden:1	8	Book:8.12.Committing:8.12.CompoundSubstance:7.38.Likely:8.75.Music:8.12.NormativeAttribute:6.5.Organism:6.12.destination:8.62.	7,72
gebonden:2	8	Book:8.12.Committing:8.12.CompoundSubstance:7.38.Likely:8.75.Music:8.12.NormativeAttribute:6.5.Organism:6.12.destination:8.62.	7,72
dreigend:4	12	Anger:9.42.Black:7.67.Clothing:9.5.Decreasing:7.5.DiseaseOrSyndrome:6.5.EmotionalState:7.75.Human:8.67.ImmediateFutureFn:8.83.Motion:5.75.Pretending:7.5.RadiatingLight:7.42.SubjectiveAssessmentA ttribute:7.33.	7,82
verstikkend:1	4	Asleep:8.Removing:8.5.SubjectiveAssessmentAttribute:6.5.causes:8.5.	7,88
onontwikkeld:3	9	Embryonic:8.11.GeopoliticalArea:6.56.HumanYouth:8.89.LandArea:7.56.Learning:8.NonFullyFormed:7.33.SubjectiveAssessmentAttribute:7.33.knows:7.89.uses:9.44.	7,9
onbeslist:1	9	Learning:7.11.Music:8.33.PsychologicalAttribute:7.33.Selecting:7.11.Stating:8.78.SubjectiveAssessmentAttribute:6.89.Unlikely:8.44.believes:7.89.inhabits:9.44.	7,93
pril:1	16	BiologicalAttribute:6.62.ChemicalDecomposition:7.75.Food:6.12.FutureFn:8.38.HumanYouth:9.Language:7.12.NonFullyFormed:8.31.OrganOrTissueProcess:9.5.OrganicObject:6.88.Plant:7.25.StateOfMind:8.31.SubjectiveAssessmentAttribute:7.88.TransportationDevice:8.62.Water:8.75.earlier:8.38.starts:9.25.	8,01
grijs:1	12	Clothing:9.ColorAttribute:6.42.Covering:9.17.Human:7.83.Icon:6.17.Organism:6.17.PastFn:9.RadiatingLight:7.5.SubjectiveAssessmentAttribute:7.67.Unhappiness:9.83.White:8.08.WhiteEthnicity:9.5.	8,03

The average mapping distance of all nouns with 4 or more mappings is 5.1, for verbs this is 4.1 and for adjectives: 5.3. It should be noted that the SUMO hierarchies for Types related to nouns (Processes, Objects, Substances) are usually more elaborated than the hierarchies for Abstract things such as adjectives. Therefore, a score of 5.3 for adjectives is much worse than a score of 5.1 for nouns.

In the next tables, we show the results for the database at April 4th for each part of speech. Scores range from -1 up to an average distance of 9 steps. If there is no intersection of two hierarchies, the score -1 is assigned. Scores are generalized to whole numbers to get a less fine-grained table. The columns represent the number of SUMO mappings: 2, 3, and 4 or more.

Table 8: Global average distance of mappings to nouns

	NOUNS					
SCORE	2 SUMO Types		3 SUMO Types		4 and more	
-1	105	1,65%		0,00%		0,00%
0	1847	29,10%	194	8,87%	48	3,31%
1	761	11,99%	201	9,19%	38	2,62%
2	916	14,43%	252	11,52%	118	8,13%
3	1371	21,60%	450	20,58%	132	9,10%
4	990	15,60%	490	22,41%	326	22,47%
5	270	4,25%	365	16,69%	259	17,85%
6	79	1,24%	150	6,86%	311	21,43%
7	9	0,14%	73	3,34%	166	11,44%
8		0,00%	12	0,55%	48	3,31%
9		0,00%		0,00%	5	0,34%
Total	6348		2187		1451	

Table 9: Global average distance of mappings to verbs

	VERBS					
SCORE	2 SUMO Types		3 SUMO Types		4 and more	
-1	18	0,86%		0,00%		0,00%
0	807	38,54%	111	13,52%	19	3,18%
1	432	20,63%	158	19,24%	30	5,02%
2	325	15,52%	179	21,80%	101	16,89%
3	284	13,56%	185	22,53%	114	19,06%
4	157	7,50%	107	13,03%	170	28,43%
5	61	2,91%	48	5,85%	84	14,05%
6	10	0,48%	31	3,78%	70	11,71%
7		0,00%	2	0,24%	9	1,51%
8		0,00%		0,00%	1	0,17%
9		0,00%		0,00%		0,00%
Total	2094		821		598	

Table 10: Global average distance of mappings to adjectives

	ADJECTIVES					
SCORE	2 SUMO Types		3 SUMO Types		4 or more	
-1	57	4,20%		0,00%		0,00%
0	216	15,91%	28	5,67%	3	0,46%
1	182	13,40%	38	7,69%	3	0,46%
2	244	17,97%	41	8,30%	32	4,86%
3	406	29,90%	132	26,72%	52	7,90%
4	210	15,46%	119	24,09%	154	23,40%
5	37	2,72%	87	17,61%	179	27,20%
6	5	0,37%	41	8,30%	192	29,18%
7	1	0,07%	8	1,62%	40	6,08%
8		0,00%		0,00%	3	0,46%
9		0,00%		0,00%		0,00%
Total	1358		494		658	

In the tables, we marked the largest proportions of synsets. For all the parts of speech, we see that these main proportions shift from low scores (low average distance) to higher scores, when more and more mappings are provided. This is what we expect. The more mappings, the less reliable the equivalence relations are and the more fuzzy and erroneous the SUMO mappings. We also see that the peak of distances for nouns and adjectives is around 5/6 for cases with 4 or more mappings. This corresponds with the averaged scores that we discussed above. Overall, these findings suggest that the average distance score is a reliable indication of wrong mappings.

The last phase of the project will concentrate on fixing the mappings for synsets with many equivalents to English wordnet and with many SUMO mappings. Especially the cases that have high distance scores will be handled. The results of this work can be measured directly by comparing the tables for the edited data with the current data.

Furthermore, we can use the scores as a quality measure for the synset to English wordnet and synset to SUMO mapping, when the database is delivered. Users of the database can use the distances score rate the properties derived through the mappings.

Finally, we expect that the measure can be further improved in two ways. First of all, we can use the disjoint in SUMO, e.g.:

(disjoint Amphibian Fish Reptile)
(disjoint AnimalLanguage HumanLanguage ComputerLanguage)
(disjoint ApartmentBuilding SingleFamilyResidence)
(disjoint Ape Monkey Hominid)
(disjoint AquaticMammal HoofedMammal Marsupial Rodent Primate)
(disjoint Article Book)
(disjoint ArtilleryCannon MilitaryTank)
(disjoint Asphalt FossilFuel)
(disjoint AstronomicalBody GeographicArea)
(disjoint Attaching Detaching)
(disjoint AutonomicProcess IntentionalProcess)

For this comparison, we need to traverse the parent tree of each class and check the compatibility of each node. Mappings with disjoint parent trees are more serious mismatches than parent trees that are not disjoint.

The second improvement is that we take the score of the equivalence relation into account. The SUMO mappings are now imported from all the equivalence relations and regardless of the mapping score of the relation. Below are for example the SUMO mappings and the equivalence relations for *godsdienstoefening* (religious service), which had an extreme number of mappings with the largest tree-distance score :

SUMO: (=, , Serving) , (+, , Repairing) , (+, , Giving) , (+, , Impacting) , (=, , SubjectiveAssessmentAttribute) , (+, , IntentionalProcess) , (=, , employs) , (+, , SexualReproduction) , (=, , ReligiousProcess) , (=, , ReligiousService) , (+, , Cooperation) , (+, , Tableware) , (+, , instrument) , (+, , EmotionalState) , (=, , ServiceOrganization) , (=, , MilitaryService)

Equivalence relations:
[EQ_NEAR_SYNONYM] service:7/ENG20-00091561-n
version: pwn_1_5 score: 2050.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] overhaul:1, inspection and repair:1, service:8/ENG20-00255586-n
version: pwn_1_5 score: 1769.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] service:11, serving:2, service of process:1/ENG20-00304340-n
version: pwn_1_5 score: 1621.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] serve:1, service:10/ENG20-00539119-n
version: pwn_1_5 score: 2221.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] service:1/ENG20-00548352-n
version: pwn_1_5 score: 2221.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] service:15/ENG20-00550265-n
version: pwn_1_5 score: 1880.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] service:5/ENG20-00555444-n
version: pwn_1_5 score: 2221.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] servicing:1, service:14/ENG20-00802650-n
version: pwn_1_5 score: 1317.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] religious ceremony:1, religious ritual:1/ENG20-00968495-n
version: pwn_1_5 score: 1832.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] worship:1/ENG20-00969068-n
version: pwn_1_5 score: 2821.0source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] rite:1, religious rite:1/ENG20-00969819-n
version: pwn_1_5 score: 1832.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] service:3, religious service:1, divine service:1/ENG20-00972198-n
version: pwn_1_5 score: 3226.0source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] service:4/ENG20-01142375-n
version: pwn_1_5 score: 2050.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] service:9, table service:1/ENG20-04016983-n
version: pwn_1_5 score: 549.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] avail:1, help:3, service:13/ENG20-04860653-n
version: pwn_1_5 score: 767.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] worship:2, adoration:1/ENG20-07041774-n
version: pwn_1_5 score: 1597.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] worship:2, adoration:1/ENG20-07041774-n
version: pwn_1_5 score: 1586.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] service:2/ENG20-07689237-n
version: pwn_1_5 score: 682.0 source_id: HEURISTICS_BI

[EQ_NEAR_SYNONYM] military service:1, armed service:1, service:6/ENG20-07700991-n
version: pwn_1_5 score: 682.0 source_id: HEURISTICS_BI

The highest scores are 3226.0 for service:3, religious service:1, divine service:1 and 2821.0 for worship:1, which are both the correct translations. We can restrict the translations and mappings or we can rank the mappings on the basis of the equivalence scores and likewise exclude incompatible mappings with lower scores.

8 Conclusions

In this deliverable, we described the design of the mapping of synsets to an independent ontology. We motivated our choice for SUMO/MILO and we explained how complex mappings can be made using relation triplets in Cornetto. We also explained how these triplets can be expanded to more formal statements, although the full expressive power of SUMO-KIF cannot be achieved. A separate discussion was held on extending SUMO to represent the meanings of adjectives. Finally, we explained how the SUMO mappings are used to measure consistency and direct further work of editing. For this we derived a mapping consistency score. This score can also be used to mark the quality of mappings in the final result.

9 References

Adjectives in Germanet, http://www.sfs.nphil.uni -tuebingen.de/lsd/Adj.html.

Atserias, J, Álvez, J., Carrera, J., Climent, S, Oliver, A, and Rigau, G., 2008, Consistent annotation of EuroWordNet Top Concept Ontology, In proceedings of the Fourth International GlobalWordNet Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008

Copestake A. and Briscoe, T (1991). Lexical operations in a unification-based framework. In: Pustejovsky J. and Bergler S. (eds.), Lexical Semantics and Knowledge Representation Association for Computational Linguistics.

Copestake, A. (1995). Representing Lexical Polysemy, in: Proceedings of AAAI Stanford Spring Symposium, Stanford 1995

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

Elkateb. S., Black, W., Farwell, D., Rodroguez, H., Alkhalifa, M., Pease, A., Vossen, P., and Fellbaum, C. (2008) *"Challenges of Building an Arabic WordNet"* in: Proceedings of the Sixth International Conference on Language Engineering , Cairo, Egypt, March 27-28, 2008.

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

Fellbaum C., M. Alkhalifa, W. Black, S. Elkateb, A. Pease, H. Rodriguez, P. Vossen, (2006) "*Building a WordNet for Arabic*", In: Proceedings of the the 5th Conference on Language Resources and Evaluation LREC2006, Genoa, Italy, May 22-28, 2006

Fellbaum C., and Vossen, P., (2007) "Universals and Idiosyncracies in Multilingual WordNets", In: Handbook Multilingual Lexicography, Oxford University Press, 2007

Guarino, N. and C. Welty (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.

Gruber, T.R. (1993) A translation approach to portable ontologies. Knowledge Acquisition, 199-220.

Horák, A., Pala, P., Rambousek, A., Povolný, M. (2006). DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. Proceedings of the Third International Wordnet Conference (GWC-06), Jeju Island, Korea.

Horák, A., I. Maks, A. Rambousek, R. Segers, H. van der Vliet, P. Vossen (2008, fc) "Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology", in: Proceedings of the 18th International Congress of Linguists (CIL18), Seoul, Republic of Korea, July 21-26, 2008.

Huang, Chu-Ren, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari and Laurent Prévot Eds. (to appear). Ontologies and the Lexicon. *Cambridge Studies in Natural Language Processing.* Cambridge: Cambridge University Press.

Hundschnurscher, F. & J. Splett (1982) Semantik der Adjektive im Deutschen: Analyse der semantischen Relationen. Westdeutscher Verlag.

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

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

Mascardi, V., Cordì, V., Rosso, P. (2008), A Comparison of Upper Ontologies (Technical Report DISI-TR-06-21), Dipartimento di Informatica e Scienze dell'Informazione (DISI), Università degli Studi di Genova, Via Dodecaneso 35, 16146, Genova, Italy

Mendes, S. (2006) Adjectives in Wordnet.Pt, In: Proceedings of GWC-2006.

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.: Linking Lexicons and Ontologies (2003) Mapping WordNet to the Suggested Upper Merged Ontology. 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.

Oberle, D., Cimiano, P., Michael Sintek,M., Kiesel,M., Mougouie, B., Vembu, S., Baumann, S., Romanelli, M., Buitelaar, P., Engel, R., Sonntag, D., Reithinger, N. Loos, B., Porzel, R., Zorn, HP, Micelli, V., Schmidt, C., Weiten, M., Burkhardt, F., Zhou, J. (2007) DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology).

Pease, A., C. Fellbaum, P. Vossen, (2008fc) "Building the Global WordNet Grid", Proceedings of the 18th International Congress of Linguists (CIL18), Seoul, Republic of Korea, July 21-26, 2008.

Peters, I. & W. Peters (2000) The treatment of adjectives in SIMPLE: Theoretical Observations, In: Proceedings of LREC 2000.

Pustejovsky, J. (1995). The Generative Lexicon, MIT Press, Cambridge MA, U.S.A.

Raskin, V. And Nirenburg, S. (1995) Lexical Semantics of Adjectives: a Microtheory of Adjectival Meaning. MCCS report 95-288.

Tufis , Dan, Verginica Barbu Mititelu, Luigi Bozianu, Catalin Mihaila, Romanian WordNet: New Developments and Applications. Proceedings of the 3rd Global Wordnet Conference, Jeju Island, Korea, South Jeju, Januaru 22-26, 2006

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

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

Vossen. P., and C. Fellbaum, (2007) "Individuals and Stuff: A Crosslinguistic Perspective", In: Proceedings of The First International Workshop on Intercultural Collaboration (IWIC 2007), Kyoto, Japan, January 25-26, 2007

Vossen P., Maks I., Segers R., VanderVliet H., van Zutphen H. (2008a) "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., Maks I., Segers R., VanderVliet H. (2008b, fc) "Integrating lexical units, synsets and ontology in the Cornetto Database", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30 May 2008.

Vossen P., E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon (2008c, fc) *"KYOTO:* A system for Mining, Structuring and Distributing Knowledge Across Languages and Cultures", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008.