Adjectives in the Dutch semantic lexical database CORNETTO

Isa Maks, Piek Vossen, Roxane Segers, Hennie van der Vliet

Faculteit der Letteren, Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV, Amsterdam
E-mail: {e.maks, p.vossen, rh.segers, hd.vandervliet}@let.vu.nl

Abstract

The goal of this paper is to describe how adjectives are encoded in Cornetto, a semantic lexical database for Dutch. Cornetto combines two existing lexical resources with different semantic organisation, i.e. Dutch Wordnet (DWN) with a synset organisation and Referentie Bestand Nederlands (RBN) with an organisation in Lexical Units. Both resources will be aligned and mapped on the formal ontology SUMO. In this paper, we will first present details of the description of adjectives in each of the the two resources. We will then address the problems that are encountered during alignment to the SUMO ontology which are greatly due to the fact that SUMO has never been tested for its adequacy with respect to adjectives. We contrasted SUMO with an existing semantic classification which resulted in a further refined and extended SUMO geared for the description of adjectives.

1. Introduction

The Cornetto project (STE05039) is funded by the Nederlandse Taalunie within the STEVIN framework. The goal of Cornetto is to build a lexical semantic database for Dutch, covering 40K entries, including the most generic and central part of the language. It will contain vertical and horizontal semantic relations as well as combinatorial lexical constraints such as multiword expressions, idioms and collocations, lexical functions and frames. The semantic layer will be validated with the formal ontology SUMO, to make it usable in Semantic Web environments. The content and semantic structure will be derived from the combination and alignment of two existing semantic resources for Dutch: the Dutch Wordnet (DWN; Vossen 1998) and the Referentie Bestand Nederlands (RBN; Maks et al 1999). RBN contains 45K entries with corpus-based descriptions, covering morphology, syntax, combinatorics, semantics and pragmatics.

2. Encoding of Adjectives

An important issue in the Cornetto project is the alignment of the word meanings in RBN and the word meanings in DWN. The database represents different approaches to semantic organisation: RBN is organised in Lexical Units, i.e. word-meaning combinations that represent a lexical form and a single meaning of this form (Cruse, 1986); in DWN, like in other wordnets, each sense of an entry is defined and determined by its synset, i.e. its set of synonyms. RBN's Lexical units contain morphosyntactic, syntactic, semantic and combinatorical information. Together, these constitute the criteria for semantic discrimination in case of polysemy. In contrast, DWN represents word meanings as concepts that are defined by lexical semantic relations to other word meanings. Here, criteria for word sense discrimination are based on these lexical semantic relations.

In the following sections, an overview will be presented of the information about adjectives that is contained in RBN (2.1) and in DWN (2.2) and that is used for semantic discrimination.

2.1 Encoding of adjectives in RBN

RBN adjectives are encoded along the following syntactic and semantic criteria:
Syntactic complementation such as oblique or fixed PP complementation, for example:

gek 1. [+ 'op'] (having a strong preference or liking for) *fond (of)*
gek 2. (affected with insanity) *mad*

Gradability and possibility of co-occurrence with degree verbs, for example:

burgerlijk 1. [non-gradable] (related to citizens) het burgerlijk huwelijk *the civil marriage*
burgerlijk 2 .[comparative: burgerlijker, superlative: burgerlijkst] (narrow-minded) burgerlijke opvattingen *parochial views*

Occurrence in attributive or predicative contexts, for example:

kapot 1.[attributive/predicative] (physically and forcibly separated into pieces) het kapotte kopje *the broken cup*
kapot 2. [only predicative] (very tired) ik ben kapot na deze lange dag *I'm beat after this long day*

Classification into a global semantic typology. Adjectives are divided into seven semantic classes: attributes of abstract nouns, emotional and mental attributes, physical and perceptional attributes, colour, substance, place, and temporal attributes. These classes refer to the relation between the adjective and the modified noun, for example:

kil 1. [physical/perception] (*disagreeably cold, chilly*)
kil 2. [emotional/mental] (*lacking warmth of feeling*, *chilly*)

Application of semantic type-shift rules which reduce the number of senses of polysemous words. These are applied to stop further sense subdivision, for example:

boos 1. [emotional/mental > abstract] (angry) een boze

vrouw (*an angry woman*), boze ogen (*angry eyes*), een boze brief (*an angry letter*)

The example illustrates the type-shift rule for the regular polysymy of emotional attributes, i.e. adjectives denoting an emotional state like boos (angry). The shift applies to the noun modified by the adjective, which may refer to a person who feels the emotion, to a bodypart, or to an object that expresses this emotion. Instead of defining a new sense, the two senses are taken together and are accounted for by using a type-shift rule.

The examples show that the criteria presented above help to disambiguate polysemous adjectives. However, they are not sufficient to distinguish all word senses. Further subdivision in RBN is based on the assessment of corpus examples and on the lexicographer's intuition.

2.2 Encoding of adjectives in Dutch WN

The adjectives in DWN are encoded according to the following characteristics:

Antonym relations
As in all other wordnets, the antonymy relation is considered as an important semantic relation. Each synset has at least one antonym relation with another synset.

Pertainyms
We distinguish between descriptive and relational adjectives. For the latter group we encode the relation which points to the noun synset to which adjectives pertain. For example:

[**natuurkundig, fysisch** (*physical*)]
pertains_to
[**natuurkunde, fysica** (*physics*)]

Hypernyms and hyponyms
Only few adjectives are organised into hypernym clusters. We applied small and flat hierarchies in the case of intensifying adjectives only. For example:

[**knotsgek, stapelgek, krankjorum, knettergek** (*very mad*)]
hyponym of
[**gek, dwaas**, … (*mad*)]

Near-synonym relations
The most productive relation is that of the near-synonym. Instead of creating large and fuzzy synsets with partial synonyms, we tried to create small synsets with complete synonyms which are related to each other by near-synonym relations. The result is a network of small closely related synsets. Consider for example the following synset:

[**dol, gek, dwaas, gaga** (*mad, crazy, foolish*) **achterlijk, gestoord** (*retarded, disturbed*)]

The involved synonyms are semantically close but not quite synonymous. We split up the synsets in two new synsets and related them with a near-synonym link:

[**dol, gek, dwaas, gaga**] 'behaving irrational'
NEAR_SYNONYM
[**gestoord, achterlijk**] 'affected with insanity'.

X-POS relations
Cross POS relations are encoded to relate lexical items that refer to semantically close related-concepts from a different POS.
[waar] (*true*) X_POS [**waarheid]** (*truth*)

Other semantic relations like STATE-OF which relates the adjective with a typical noun, are encoded only for few cases. For instance:
[glazig] (*waxy, soapy*)
STATE_OF
[**aardappel**] (*potato)*

2.3 Adjectives in Cornetto

Cornetto combines the semantic structures of RBN's Lexical Units and DWN's synsets by automatic alignment, followed by manual editing. Automatic alignment performed relatively poorly in case of adjectives, since each resource has its own system of defining word senses, focusing on different distinctions. To address this problem, a further classification of adjectives was needed. 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 selected the 100 most frequent and also most problematic adjectives, - i.e. words with many-to-many relations between the RBN and DWN senses - and succeeded in aligning them by redefining synsets and LUs using this more refined classification. The following example shows how the H&S classification helped to resolve an alignment problem between RBN and DWN.

In RBN the adjective *kort (short)* is monosemous;
in DWN however, a spatial and a temporal sense are distinguished:

RBN:
Kort (*short*) 1. [of time and length]
Collocations: een korte dag (*a short day*); een korte vakantie (*a short holiday*); een korte broek (*short trousers*) ; kort haar (*short hair*).
DWN:
[kort] antonym [**lang:1]** (*long*)
[**kort** (*short*)**, kortdurend** *(short-lived)*] antonym [**lang:2**] (*long*)
The classification supports the second approach as it has different classes for time-related and size-related adjectives. The LU in RBN is split up in two different LUs

which can now be aligned to the DWN synsets; the DWN synsets are completed with further semantic relations. In this case, we can solve in a rational way a problematic alignment.

3. Mapping adjective synsets to SUMO

The final goal is to map the synsets to the SUMO ontology. The ontology is seen as an independent anchoring of meaning across languages. Furthermore, it is a more formal way of representing meaning that can be used by machines for inferencing. Certain semantic implications are made explicit in the ontology and not in the lexicon, whereas other more linguistic data are not presented in the ontology.

The mapping of the Dutch synsets to SUMO is copied from the English Wordnet. In the English Wordnet-to-SUMO mapping, each synset is related to a single SUMO term, mostly by an equivalence relation (=) or a subsumption relation (+). The adjectival synsets in DWN are automatically mapped to adjectival synsets in the English Wordnet. Through this relation, SUMO Terms are assigned to the adjectival synsets. However, the automatic English Wordnet to SUMO mappings are not yet corrected manually as far as the adjectives are concerned. This has two major implications: (1) the number of incorrect default mappings is high (2) it has not yet been checked if the ontology is complete. By consequence our task is twofold: we need to revise most adjective mappings and we need to check and revise SUMO in order to achieve the coverage of all concepts that are expressed by general language adjectives.

3.1 SUMO for CORNETTO

All attribute classes are subsumed by either RelationalAtribute or InternalAttribute. 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 (e.g. *wide*)
- LengthAttribute (e.g. *short*)
- BiologicalAttribute (e.g. *hungry*)
- PhysicalAttribute (e.g. *dense, wet*)
- PerceptualAttribute (e.g. *loud, aromatic, sour, rough*)
- 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 the H&S classification. Additionally, we tried to map the 100 most-frequent adjectives (which we already had mapped on the H&S classification) with the SUMO as well. When comparing the H&S classification with SUMO, many main classes could be easily transferred (see Table 1)[1].

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

Table 1 comparison of H&S semantic classes and SUMO terms

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.

- 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 PsychologicalAttribute. In this particular case we introduced the class BodilyAttribute which subsumes all kinds of attributes referring to the human and 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:

[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.

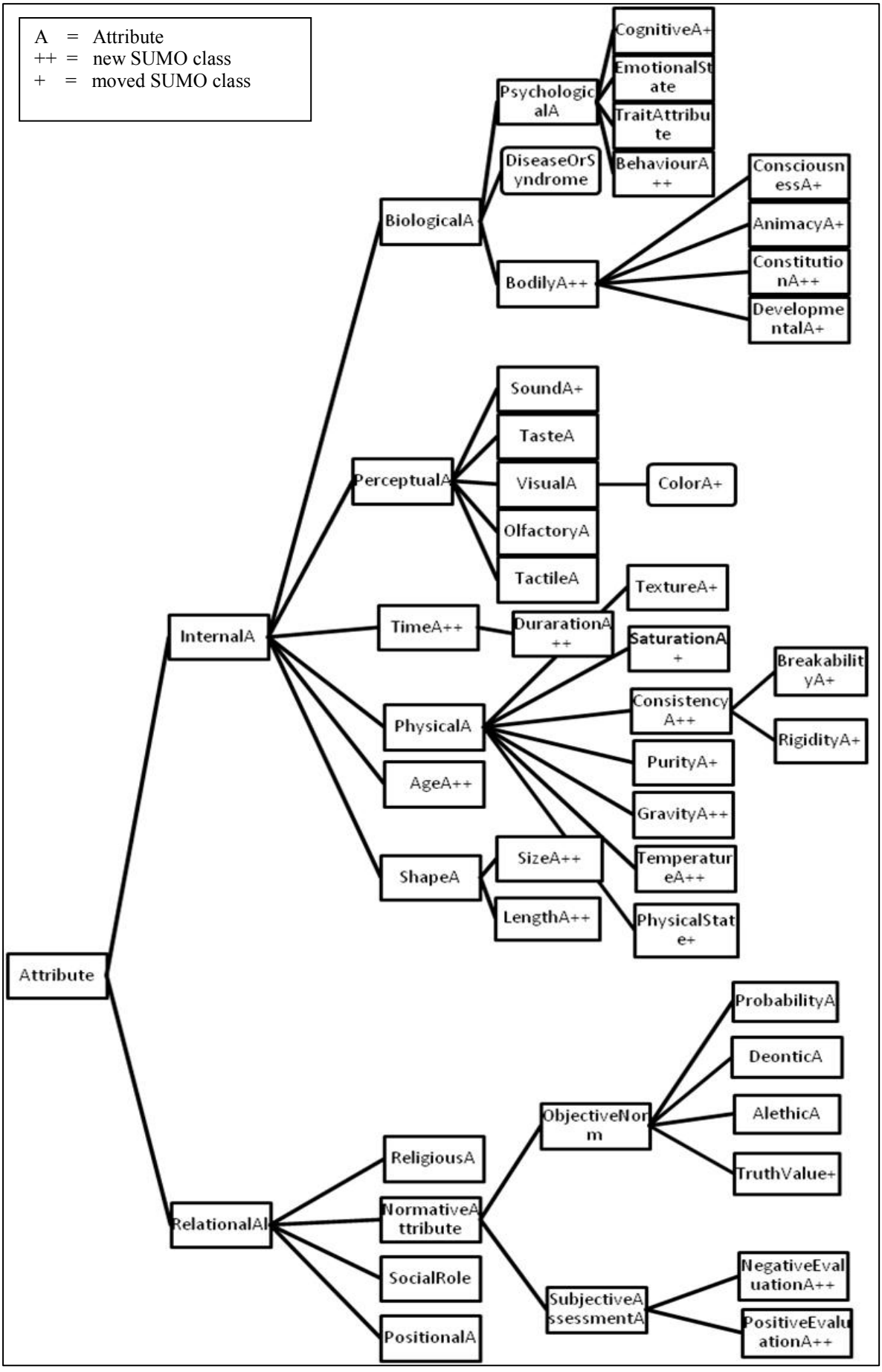

Figure 1 SUMO for Cornetto adjective concepts

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, and "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 colour, 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 colour. 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 colour. Thus, a property that is seemingly truly 'internal' such as colour, may be considered relational as well.
The example shows that contrasting the two systems requires resolution of certain questions that are not completely unambiguous. In a practical sense this poses a problem, but 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.

4. Conclusions and Future Prospects

We described the merge of the SUMO ontology with an existing semantic classification of adjectives to make the former more complete with regard to the coverage of general-language adjectives. The resulting ontology is a starting point for future work on a more detailed semantic representation by using SUMO and SUMO axioms.
The work on Cornetto is still ongoing and will be completed by 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

5. Acknowledgements

This research has been funded by the Netherlands Organisation for Scientific Research (NWO) via the STEVIN program for stimulating language and speech technology in Flanders and The Netherlands.

6. References

Adjectives in Germanet, http://www.sfs.nphil.uni-tuebingen.de/lsd/Adj.html.
Cruse, D. (1986) Lexical semantics. University Press, Cambridge.
Fellbaum, C. (ed.) (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge MA.
Hundschnurscher, F. & J. Splett (1982) Semantik der Adjektive im Deutschen: Analyse der semantischen Relationen. Westdeutscher Verlag.
Maks, I., Martin, W., Meerseman, H. de (1999) RBN Manual, Vrije Universiteit Amsterdam.
Mendes, S. (2006) Adjectives in Wordnet.Pt, In: Proceedings of GWC-2006.
Niles, I., Pease, A. (2001) Towards a Standard Upper Ontology. In: Proceedings of FOIS 2, Maine.
Peters, I. & W. Peters (2000) The treatment of adjectives in SIMPLE: Theoretical Observations, In: Proceedings of LREC 2000.
Raskin, V. And Nirenburg, S. (1995) Lexical Semantics of Adjectives: a Microtheory of Adjectival Meaning. MCCS report 95-288.
Vliet, H.D. van der (2007) The Referentie Bestand Nederlands as a multi-purpose lexical database. In: International Journal of Lexicography 20.3.
Vossen, P. (ed.) (1998) EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer, Dordrecht.