CATEGORIES AND CLASSIFICATIONS IN EUROWORDNET

Piek Vossen and Laura Bloksma

University of Amsterdam

1012VB, Amsterdam, The Netherlands

[piek.vossen@let.uva.nl; lbloks@let.uva.nl]

Abstract

In EuroWordNet we develop wordnets in 8 European languages. which are structured along the same lines as the Princeton WordNet. The wordnets are inter-linked in a multilingual database, where they can be compared. This comparison reveals many different lexicalizations of classes across the languages that also lead to important differences in the hierarchical structure of the wordnets. It is not feasible to include all these classes (the superset) in each language-specific wordnet and to reach consensus on the implicational effects across all the languages. Each wordnet is therefore limited to the lexicalized words and expressions of a language. The wordnets are thus autonomous language-specific structures that capture valuable information about the lexicalization of each language, which is important for information retrieval, machine translation and language generation. By connecting the wordnets to a separate ontology. semantic inferencing can still be guaranteed. Still, different types of classification schemes can be distinguished among the lexicalized classes. In this paper we will further describe the properties of these different classes and discuss the advantages and effects of distinguishing them in wordnet-like structures.

Introduction

The aim of EuroWordNet is to develop a multilingual database with wornets in 8 European languages: English, Dutch, Italian, Spanish, German, French, Czech and Estonian. Each language-specific wordnet is structured along the same lines as WordNet (Miller *et al.*, 1990): synonyms are grouped in synsets, which in their turn are related by means of basic semantic relations such as hyponymy (between specific and more general concepts), meronymy relations (between parts and wholes). By means of these relations all meanings can be interconnected, constituting a huge network or wordnet. Such a wordnet can be used for making semantic inferences about the meanings of words, for finding alternative expressions or wordings, or for simply expanding words to sets of semantically related or close words in information retrieval. Furthermore, semantic networks give information on the lexicalization patterns of languages, on the conceptual density of areas of the vocabulary and on the distribution of semantic distinctions or relations over different areas of the vocabulary.

The most important difference of EuroWordNet with respect to WordNet is its multilinguality, which, however, also raises some fundamental issues with respect to the language-specificity of the semantic relations. Multilinguality is achieved by adding an equivalence relation for each synset in a language to the closest synset in

WordNet 1.5. Synsets linked to the same WordNet 1.5 synset are supposed to be equivalent or close in meaning and can then be compared. In the ideal case, we would expect that, for example, the Dutch nouns *doos* (box), *tas* (bag), *asbak* (ashtray), *lepel* (spoon) are related to the same hyperonym container as the WordNetl.5 equivalents. However, in Dutch there is no direct equivalent for *container*. 1 As a result of this we see that these concepts are directly linked below the equivalent of *object* (voorwerp) in the Dutch wordnet, see Figure 1.

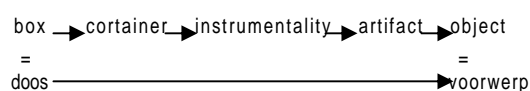Figure 1 also illustrates another difference. In addition

Figure 1: Lexicalized and non-lexicalized classes.

to the lexicalized classes, the WordNetl.5 hierarchy also includes non-lexicalized or artificial categories, such as *instrumentality, natural object, external body part,* which you will not likely find as an entry in a general dictionary of contemporary English. The Dutch wordnet, on the other hand, only contains categories lexicalized in the language, which makes the Dutch hierarchy much flatter and less rich than WordNet 1.5.

In EuroWordNet, we will be dealing with many differences in lexicalization across languages and consequently differences in the hierarchies as well. Furthermore, also the lexicalized classifications do not form a homogenous set. In addition to classical taxonomies for *plants* and *animals,* there turn out to be a gamut of words that could also be used: *vermin, pet, carnivore,* etc. Typically, the wordnets contain various mixtures of different classes and it is not clear what classificational perspective should be chosen and what the effect is for usability of the wordnets.

In this paper we will describe the different classifications and clarify their usage in different applications. In the next section, we first distinguish the role of lexicalized and non-lexicalized classes and argue for distinguishing different ontologies for capturing lexicalizations and for making inferences. After that, we further differentiate the lexicalized classes on the basis of their conventionality and circumstantiality. We will demonstrate how they can be integrated in a single wordnet hierarchy.

---

[1] The word *container* in Dutch does exist but is only used for big containers on ships or for big garbage cans.

Lexicalized and non-lexicalized categories

Artificial categories, such as *instrumentality, external body part, plant parts,* clearly help to group related concepts that share a meaning component and give more structure to the hierarchy. A deeper and richer structure makes it possible to infer more properties for concepts, i.e. it is not possible to derive properties, such as *containing, artificial* and *functionality,* from the hyponymic relations for the Dutch concepts in Figure 1. The disadvantage of an 'artificial' hierarchy as WordNetl.5 is that it does not give correct predictions about the substitutability of the nouns: e.g. a speaker of English will not use the noun *instrumentality* to refer to *containers, boxes, spoons,* and *bags.* This is particularly relevant for information retrieval systems or language generation modules that have to deal with lexical choice.

Another problem, which is specifically relevant for EuroWordNet, is that there is no a priori reason why we should limit ourselves to the WordNetl.5 classes only. We may as well take over all lexicalized classifications occurring in all the 8 wordnets, giving us a more universal set. We could also continue to invent new classes and expressions to capture more and more generalisations, and we may end up with adding any conceivable semantic property as a class to create very rich inheritance structures. Apparently, such a strategy has been followed for artificial ontologies, such as Cyc (Lenat and Guah, 1990), which are purely designed for structuring knowledge. Classes, such as *AnimalBodyPart, ContainerProduct, SolidTangibleThing, SomethingExisting,* are not intended as lexical entries in a lexicon in the first place.

In all these options, the wordnets are no longer networks of lexicalized words and expressions in languages. Still, they do not automatically give us a good procedure for building a conceptual ontology to inherit properties. There are many different ways in which the same knowledge can be stored (Gruber, 1992). For example, in the case of the Dutch wordnet, it is possible to express the role as a *container* by a separate relation to the verb *bevatten* (to contain) for each of the objects that have such a function. The same properties can also be expressed by different orders of classification. We can first classify concepts in terms of their constitution, *substance* or *object,* and then in terms of their function, *food* or *container,* but we can also structure the information by determining the function first and then the constitution. Alternatively, we can avoid an explosion of levels by allowing multiple hyperonyms, e.g. to both *container* and *artifact* or *container* and *natural object* in WordNetl.5, creating a tangled hierarchy instead of a tree. In all these cases, the effect for inheriting properties would be the same.

Apparently, there are two different purposes for wordnets that do not always combine: making semantic inferences and capturing lexicalization patterns for substitution (Vossen, 1995). In EuroWordNet, it is certainly not feasible to develop a universal ontology for all the languages that also captures lexicalization differences and predicts substitution of words and expressions. This would imply that we reach consensus on all concepts, relations and implications across all the languages and cultures. Unless we introduce 'artificial classes' that represent the union of all classifications occurring in all the languages, it is clear that it will be unavoidable that the wordnets will exhibit important structural differences in the hierarchies. We therefore take the position that the wordnets only and exactly reflect the lexicalization patterns in a substitution network. Each wordnet should thus be seen as an autonomous language-specific structure. The wordnets are lexical ontologies rather than conceptual ontologies. In a conceptual ontology ft may be necessary to introduce artificial non-lexicalized levels to structure the knowledge or it may be necessary to neglect lexicalized levels which are not relevant for the purpose of the ontology (e.g. *waste product, threat, favorite).* A lexical ontology, on the other hand, may not have particular classes that entail important properties and it must also include many classes that are not relevant for structuring knowledge.

In the EuroWordNet database, the autonomous wordnets are inter-linked via an unstructured Inter-Lingual-Index. The only purpose of the Inter-Lingual-Index is to provide an efficient matching across the wordnets (Vossen *et al,* 1997, Peters *et al.,* fc). In addition, the database can still be extended with an ontology or knowledge base that is structured for the purpose of making inferences. By connecting this ontology to the same Inter-Lingual-Index, all the wordnets will get access to the properties stored in the ontology (see Figure 2). Currently, the wordnets are connected to the EuroWordNet top-ontology. This ontology will be further extended with the Reference Ontology that is being developed by the ANSI Committee for Ontology Standards. These ontologies can be structured and organized along very different principles and structures (Guarino, 1997, Rodriquez et al, fc, Sowa, fc.).[2]

The wordnets provide the mapping of the language-specific lexicalizations on the shared knowledge-structures, and at the same time maintain information on the substitutability of words in the language via the language-internal relations. In Figure 2, this is illustrated for the concept BOX, which is defined in the knowledge base by a combination of classifications, but is linked to the same Inter-Lingual-Index as the meanings in the wordnets, which exhibit a separate network of language-internal relations. The Reference Ontology classes (ContainerThing and SolidTangibleThing) are derived from the public part of Cyc, combined with Pangloss, and Sensus (Hovy, fc). The EuroWordNet Top Ontology classes (Form, Function, Composition and Origin) are based on the Qualia-structure in the Generative Lexicon (Pustejovsky, 1995).

---

[2] The EuroWordNet top-ontology consists of 63 semantic features that form a partial lattice from which combinations of features can be combined. In total 450 feature combinations have been derived to classify 1024 most-important Inter-Lingual-Index concepts.
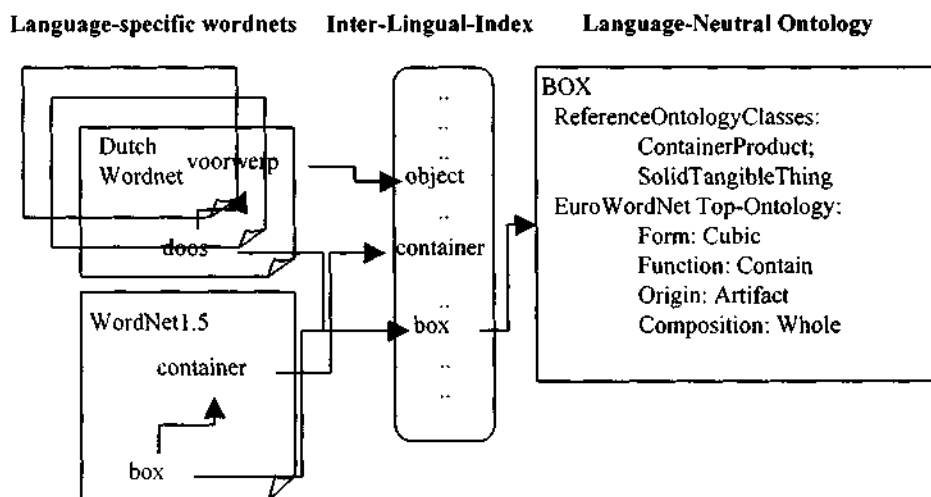
Figure 2: The integration of language-specific wordnets and language-neutral ontologies in EuroWordNet.

Types of Lexicalized Classifications

The main purpose of the wordnets is thus to reflect lexicalization and predict substitution. These patterns can be encoded in the form of lexical semantic relations such as hyponymy and synonymy between words. A possible instrument to decide on these relations is a *Diagnostic Frame* (Cruse, 1986): it is a *dog* therefore it is an *animal; *it is an *animal* therefore it is a *dog.* According to Cruse, this test reveals hyponymic and synonymous subsumption relations. Another useful test is the Principle of Economy (Dik, 1978), which states that a word should not be defined in terms of more general words if there are more specific words that can classify it:

> If a word Wl *(animal)* is the hyperonym of
> W2 *(mammal)* and W2 is the hyperonym of
> W3 *(dog)* then W3 should not be linked to Wl
> but to W2.

This principle should prevent that senses are linked too high up in the hierarchy and that intermediate levels are skipped.

What this procedure does not give you is however the set of words or classes to which we should apply the test. Intuitively, you may think of classes such as *animal, object, substance,* but there is no guarantee that we have considered all the possible classes. This immediately becomes clear when different wordnets are compared on a larger scale and a wide variety oft classification schemes are revealed.[3] So far, we distinguished the following types:

1. conventional classes: *substance, artifact, object, animal*

2. specialized classes: *vertebrate, chemical compound, mineral*

3. circumstantial classes: *waste product, favorite, threat, material*

The difference between a conventional classification and a specialized classification is shown in Figure 3. On the left is the hierarchy for *horse* in WN1.5 on the right the one for the equivalent word in the Dutch WordNet *(paard).* The difference in sublevels is due to the fact that the Dutch resource is based on contemporary/general vocabulary whereas WN 1.5 uses very specialized sublevels, only clear and useful for experts. The glosses for WN1.5 do not always make it clear how these sublevels are determined. The origin of most entries is Latin or Greek and they are labeled as zoological in the bilingual resource. More specialized lexical resources for Dutch may or may not have these sublevels; it is clear that for general vocabulary these sublevels are not appropriate.

Given the purpose of the hierarchy, information retrieval and language generation, it is useful to have all the information of all the hierarchies established, but also to maintain the difference between expert/specialised levels and levels for general/conventional vocabulary. We thus should be able to predict that *equid, hoofed mammal* and *chordate* can refer to entities of the type *horse* and that the inference is correct but also that the expectation and inference is only relevant in a certain context and for certain speakers. Miller *et al.* (1990) suggests that the familiarity of a classification scheme is reflected by the polysemy of the classes (the more common a word is used the more polysemous it is). If we restrict the classification to hyperonyms with more than one sense we get the layman variant: *horse - animal - being - entity.* All the specialised classes only have a single meaning. Polysemy is however not consistent across resources. In the Dutch wordnet *also paard en dier* have a single meaning. Still, frequency of the Dutch words according to the Celex database seems to confirm their hypothesis. The more specialised class *zoogdier* has a

---

[3] Alternatively, we could look at distribution evidence from corpora, to see which words occur in similar clusters (Church and Hanks 1990, Grefenstette 1994). However, from these clusters alone we cannot infer the precise semantic relation between the words.

```
horse 1 ————————————————————————————— paard 1
    HYPONYM_OF equid 1                          HYPONYM_OF zoogdier 1
        HYPONYM_OF odd-toed ungulate 1         HYPONYM_OF dier 1
            HYPONYM_OF hoofed mammal 1         HYPONYM_OF organisme 2
                HYPONYM_OF eutherian 1         HYPONYM_OF wezen 1
                    HYPONYM_OF mammal 1          HYPONYM_OF entiteit 1
                        HYPONYM_OF craniate 1
                            HYPONYM_OF chordate 1
                                HYPONYM_OF animal 1
                                    HYPONYM_OF being 1
                                        HYPONYM_OF entity 1
```
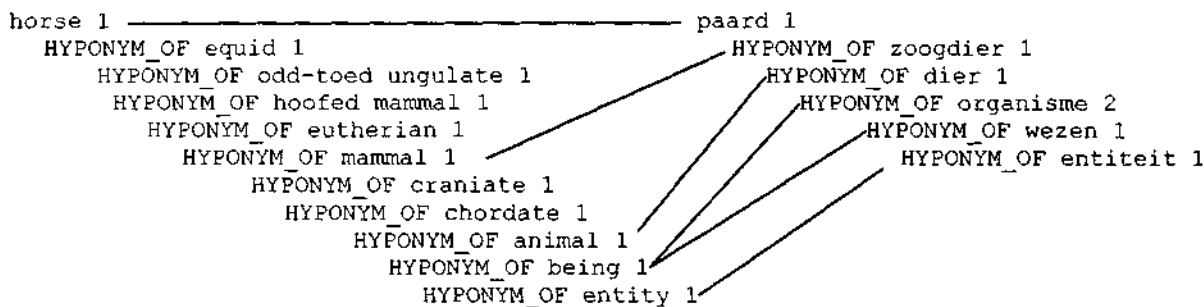
Figure 3: Specialised classification in WordNet and Conventional classification in the Dutch wordnet

frequency of 240, which is significantly lower than 6675 for paard *(horse)* and 7772 for dier *(animal).*

However, frequency and familiarity alone are not sufficient. Another important characteristic is namely the uniformity of the domain: if *eutherian* applies then all the other biological classes in the hierarchy apply as a single coherent micro-theory. If two micro-theories from different domains are intermingled, the low frequency/polysemy of the specialised classes will not differentiate the expectations. It is therefore necessary to use a kind of domain labelling in addition to the familiarity. A biological classification would automatically select all other classifications within the same coherent scheme, but not a classification in e.g. the domain of *horse racing.*

We thus propose that the hyperonyms of all domains or micro-theories are encoded together with the conventional common-sense hyperonyms in a single unified tangled hierarchy. Domain information should be encoded for each class: explicitly, in the form of an ontology of domain labels (e.g. *science,* subdivided into *medicine; physics', linguistics,* etc., *sports,* subdivided into *water sports, winter games, ball games,* etc.) and in the form of distribution information from corpora (Church and Hanks 1990, Grefenstette 1994). In the latter case, we encode for each word the likeliness it will co-occur with any of its hyperonyms in a single document. We can then make the following predictions:

- low frequent classes of the same domain will show a bi-directional consistent or mutual co-occurrence distribution, *eutherian* will strongly co-occur with *vertebrate* and the other way around;
- low frequent classes of different domains will not show any significant co-occurrence distribution;
- high frequency of conventional classifications necessarily implies a non-selective co-occurrence distribution. Co-occurrences with specialised classes are therefore not bi-directional or mutual. The specialised class *eutherian* may co-occur with *animal* but *animal* will not significantly co-occur with *eutherian,* not more than with any other candidate.

The co-occurrence correlation between two words can be obtained by dividing the frequency that a word occurs with the other word in a single document (or a corpus of a single domain) by the total frequency of

word in a large diverse corpus that is not specific for a domain (Sanfilippo 1997):

$$prob_{Domain} (W1|W2) = count_{Domain} (W1|W2)$$

$$count_{LargeCorpus} (W1)$$

This will yield a high value for low frequency words with occurrences in a single document or a single domain corpus, and low values for words with very high frequencies in general, despite their frequency in the domain-specific corpus or document. Only words that have mutual high and low correlations in the same domains or documents show a consistent domain correlation. To measure the co-occurrence correlation for all words in all documents is a lot of work. However, since we only want to know the correlation for the words related by hyponymy, it is thus only necessary to measure the correlation for the direct hyponyms and hyperonyms of a word.

As suggested in Vossen *et al.* (1995), this distribution information can be used to fine-tune a semantic hierarchy to the relevant domain. If a specific document exhibits a coherent distribution pattern, only that part of the tangled hierarchy should be considered which fits the pattern. The lexical density in that part **of** the hierarchy is thus also reduced to the relevant distribution only (the concepts in the domain and the general conventional concepts). Closeness of meaning *a* then not determined by the distance in the complete hierarchy (Resnik 1995, Agirre and Rigau 1986) but **in** the reduced hierarchy: in a text about *horse racing* the biological sublevels should not be considered **to** measure the closeness between *horse* and *animal.*

Why should general and specialized hierarchies be combined in the first place? Obviously, **it** would be easier to maintain a single hierarchy instead of rebuilding the general part of it for each new application. Another aspect is that an integrated wordnet that anchors expert terminology in general vocabulary opens wider possibility to develop applications for non-expert users as well (e.g. patients who want to get access to medical documents). It is clear that such integration is not trivial, because the expert hierarchy cannot just be added at the specific levels of the generic hierarchy but creates many tangled and intermediate levels. In the case of *horse* we would thus have a short conventional chain and a much longer specialized path which may intersect with the

conventional chain at several points (see Figure 4 below).

Unfortunately, the hierarchy becomes even more complicated when we also include so-called circumstantial classes. For example, the gloss for the entry *paard* (horse) lists some alternative classifications not yet considered:

paard l   groot viervoetig, rij-, trek- en lastdier
(horse)   (big four-footed riding-, draught and pack animal)

A *horse* is an *animal,* but not necessarily a *riding-* or *draught* or *pack animal.* The diagnostic frame for hyponymy does not give a clear negative or positive result here: it is a *horse* therefore it is a *riding animal,* but the classification is not always appropriate. Some *horses* may and others may not. This is confirmed by one of the hyponyms of *paard* (horse) which is *rijpaard* (riding horse). The fact that *rijpaard* (riding horse) is already linked to the hyperonym *rijdier* (riding animal) indirectly predicts that some horses may be called *rijdier*. It is however not just a matter of classification. If we look at other hyponyms below animal in the hierarchy we see many more potential classes (some of which include hyponyms of *paard,* others do not):

rijdier (riding animal)
        rijpaard (riding horse)
                damespaard (ladies' horse)
                jachtpaard (hunter)
pakdier (pack animal)
        pakpaard, lastpaard (pack horse)
        pakezel (pack donkey)
trekdier (draught animal)
        trekhond (draught dog)
        trekpaard (draught horse)
        trekos (draught ox)
fokdier (breeder)
        fokpaard (breed horse)
        fokstier (stud bull)
        fokschaap (breeding sheep)
        fokzeug (brood sow)
afferdier (sacrificial animal)
        paaslam (lamb to sacrifice for Easter)
proefdier (animal for experiments)
        proefkonijn (literally, rabbit for experiments)
kuisdier (pet/domestic animal)

This list is not complete, but clearly shows that in a semantic field, which is traditionally thought of as a classical taxonomy, we still see a whole range or gamut of classes that may apply to *horses* in different degrees. Some classifications only apply in rather special circumstances. There is no intrinsic property that makes an *animal* an *offerdier* (animal for sacrificing) or *proefdier* (animal for experiments). These classes are called circumstantial: they can only be applied when the context allows this. They typically have a very strong implication that creates or limits the context in which they can be used

In some aspects, circumstantial classes are similar to specialized classes. Both are non-polysemous, have low frequency and are thus also limited to a specific context. But, while specialized classes often form deep consistent chains of a single domain, the circumstantial classes are isolated and not coherent at all. Typically, they represent a diverse spectrum of interests and contexts, and hardly have any hyponyms. We thus expect that the have no significant co-occurrence preference with other classes, except for the conventional class to which they are linked.

Another important feature of circumstantial classes is that they are often not limited to a particular type of thing but can be applied to a variety of things. Likewise, we find many more of these classes when we go up the hierarchy. Above *animal,* we may thus find more concepts that could be applied to *horses,* under certain, sometimes special, circumstances (e.g. *horse racing*):

> rivaal (rival); kampioen (champion); ster (star); nakomeling (descendant); bastaard (bastard); winnaar (a winner); verliezer (a loser); ontdekking (a discovery); openbaring (a revelation); bezit (a property); blikvanger (eye-catcher); zwerver (a drifter, animals that stray) afknapper (a letdown); uitschieter (an extreme); tegenvoeter (an antipode); ramp (a disaster); mislukkeling (a failure); vergissing (a mistake); doelwit (a target); lading (cargo); verzending (shipment); offer (anything sacrificed).

As explained in Vossen (1995), these extreme cases are all located high up the hierarchy and have no, or only incidentally, hyponyms. Likewise, they are often defined with void heads: *anything that, that which, something that,* of with disjunctive heads: *a person, thing or idea that, a person, machine or animal that..* Finally, they are often derived from adjectives or verbs or compounds including an adjective or verb that captures the conceptualization. The process is very productive: about 10% of the concrete nouns in a general English and Dutch dictionary fits the above description (Vossen 195).

A final important difference of circumstantial classes with conventional and specialized is that the circumstantial classes are hardly ever disjunctive. In a conventional and specialized hierarchy, classes are complementary: something either is an *animal, plant* or *person.* The fact that many circumstantials are 'open' implies that they can cross-classify any other classification or circumstantiality. This poses a problem for the hierarchy as a network that predicts substitutability. Strictly speaking, substitution is only allowed for words that are synonyms, hyperonyms and perhaps in some cases hyponyms of a certain word. Typically, you want to exclude co-hyponyms at the same level or higher levels. This prevents that we find articles on *cats, dogs, fish, insects, plants, artifacts* when we ask for *horses* in an information retrieval application that makes use of semantic networks. It would however also exclude all the circumstantial classes at the same or higher levels, because there is no hyponymy relation between *horses* and the above circumstantial classes.

A circumstantial category at a very high level of the hierarchy can substitute everything from a

classification point of view. It is thus only restricted by the circumstances that it evokes through its meaning. This cannot be capture by a hierarchy as we have discussed so far. It may be obvious that nobody really wants to classify a *horse* using any of the above circumstantial classes, let alone to cross-classify all specific meanings with all the circumstantials. To correctly predict substitutability we propose to explicitly encode which co-hyponyms are complementary, e.g. *animal, plant, person.*

Disjointness can be encoded in EuroWordNet using labels attached to the relations. The labels Disjunct and Conjunct are typically used to specify the relation between multiple hyperonyms: *spoon* is **both** *cutlery* and *container* (Conjunct) and a *female* is **either** a *person* or an *animal* (Disjunct). The default relation is non-exclusive. By applying this principle also to hyponyms, we can differentiate between exclusive and non-exclusive hyponyms (conjunctive hyponyms do not occur). If the label Disjunct is assigned to a hyponym it means that it is complementary with other hyponyms with this label. If there is no such label the hyponyms can cross-classify each other. Instead of assuming that co-hyponyms are complementary by default it is thus necessary to explicitly encode this to prevent substitution. Circumstantial categories are normally not complementary and will therefore not get a Disjunction label. This means that substitution may spread upwards to classes which are not disjoint. Obviously, the low frequency and distributional non-preference of a circumstantial class will limit the expectation of the circumstantial class but it will not be excluded as a substitute for *horse* or any other specific concept.

Most circumstantials will be classified high up the hierarchies by abstract and void heads and the above solution very well captures their implicit vagueness. For some circumstantials this may however not be sufficient. For example, it is not wrong or odd to classify a *paard* (horse) as a *rijdier* (riding animal), *pakdier* (pack animal) or *trekdier* (draught animal), a *kat* (cat) and *hond* (dog) as a *huisdier* (pet). Furthermore, it is very informative to know that the typical *rijdieren* (riding animals) are *horses* and typical *huisdieren* (pets) are *cats, dogs* and maybe a few others in Dutch. When such stereotypical class-membership applies, the class is more or less defined by the hyponyms, instead of the hyponym by the hyperonym. This reversed dependency of certain hyperonyms (Vossen 1995) can be expressed in EuroWordNet using the label *reversed* that specifies the implication direction of relations. The label reversed is typically used for relations where the implicational expectation or dependency can vary (Alonge *et al.*, fc.):

car	HAS_MERONYM	wheel	
wheel	HAS_HOLONYM	car	**+reversed**
mouse	HAS_HOLONYM	computer	
computer	HAS_MERONYM	mouse	**+reversed**
pet	HAS_HYPONYM	cat	
cat	HAS_HYPERONYM	pet	**+reversed**
cat	HAS_HYPERONYM	animal	
pet	HAS_HYPERONYM	animal	

In the case of the first meronymy example, the label reversed indicates that *cars* may have *wheels* but *wheels* are not necessarily parts of *cars*. In the case of computer and mouse this implication is the other way around. We can use the label in the same way for hyponymy. The default relation is then from hyponym to hyperonym, but for the circumstantials with stereotypical membership relations we can reverse the default using the label. The reversibility can also be encoded as a culture-specific expectation without affecting the regular conventional classification: in some countries *camels* or *elephants* will more easily be seen as riding and/or draught animals.

Figure 4 below shows how the different types of hyponymy relations can be integrated. We have only included some examples for illustration. There are many more circumstantial and expert levels that should be included. The numbers below each node indicate the probability in a large-scale diverse corpus (Celex). Different arrows represent the 4 different types of hyponymy. Below *wezen* (being) we find a number of abstract circumstantials. The arrow indicates a non-exclusive hyponymy relation, which means that the co-hyponymic classes may overlap and intersect. Three hyponyms are disjunctive: *persoon* (person), *dier* (animal) and *plant* (plant). This means that referents belong to either one of these. They may still overlap with the non-exclusive co-hyponyms.

Below animal we find also a combination of circumstantial and disjoint classes: *beestje* (a pet name for an animal), *huisdier* (pet), *rijdier* (riding animal), *lastdier* (pack animal) and *fokdier* (breeder) as circumstantials; *paard* (horse), *kat* (cat), *hond* (dog) as disjoint, exclusive classes. We also see some cases of reversed hyponymy where there is an addition stereotypical link from certain circumstantials to *paard* (horse), *kat* (cat), and *hond* (dog). Finally, at the right side, there is a separate extension with a few multiple levels for the biological expert classification.

At the third level, we see specific circumstantials limited to horses and two complementary and therefore disjoint hyponyms *hengst* (stallion) and *merrie* (merry). According to our definition, we can now expect that:

- substitution and key word expanding may include the hyperonyms and all co-hyponyms which are non-exclusive;
- reversed hyponymy activates downward expansion to more specific concepts;
- if one expert classification applies, all expert classifications will apply;

If a user types a query with *paard* (horse) we can not only expand to the usual hyperonyms *dier* (animal) and *wezen* (being) but also to: *beestje; rijdier; lastdier; huisdier,* and even further (but weaker) to *ouder; bastaard; rivaal; nakomeling; kampioen; ster.* Furthermore, when a user types *rijdier* (riding animal), we will also get a downward expansion to *paard* (horse) and any other typically *riding animal* that is linked to it.

Finally, the hyperonym chain for the expert domain is only relevant when the domain is also

activated. Conceptual distance measurement (Resnik 1995, Agirre and Rigau 1986) will thus give a different result when the expert differentiation is only included when relevant.

normally only refer to humans but are used here to refer to *horses: kind* (child), *vader* (father), *moeder* (mother), *zoon* (son).
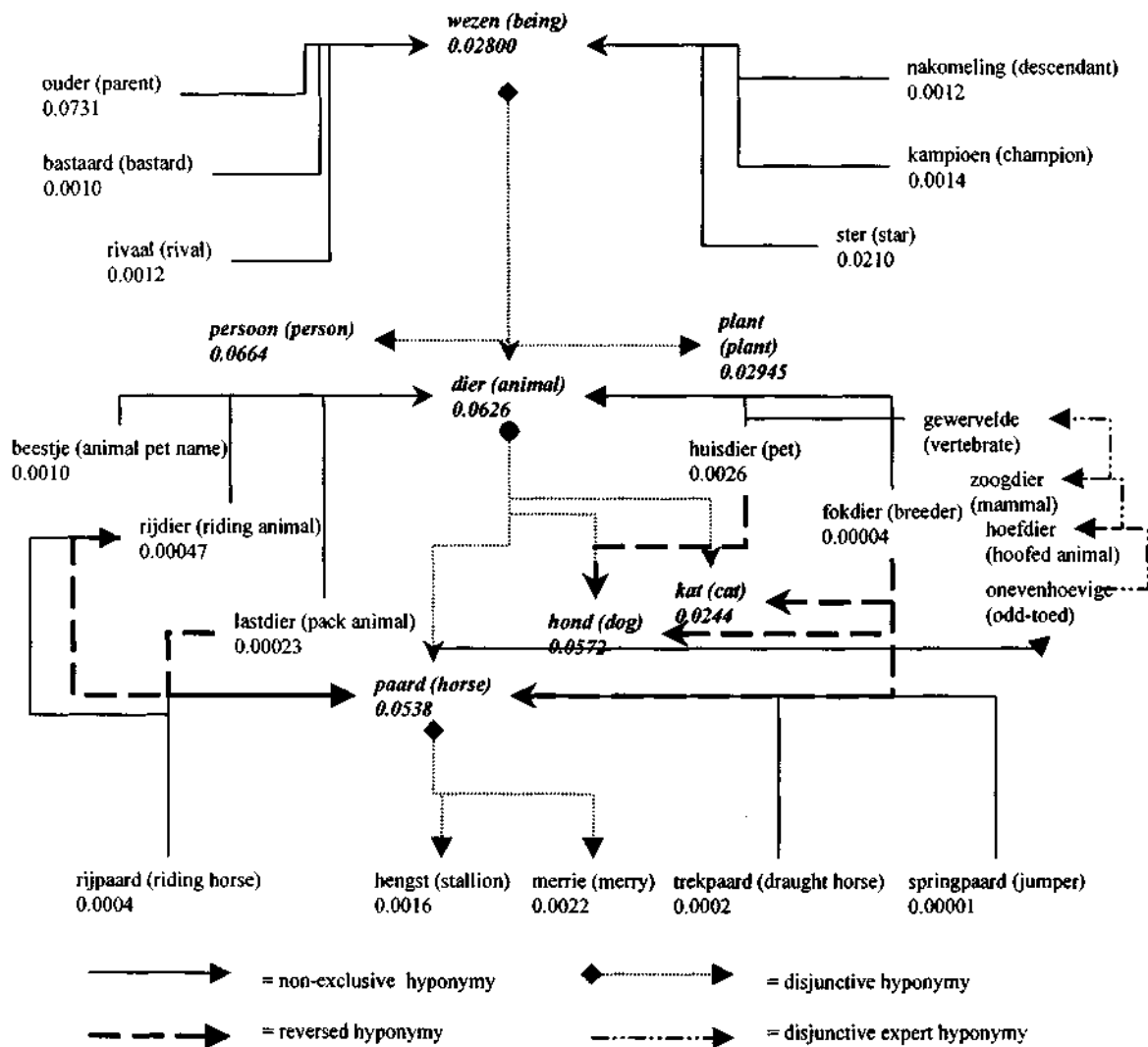

Figure 4: Integrated Hierarchy with different types of hyponymy

Domain Probability Experiment

To test the validity of the above claims, we are setting up experiments to measure the correlation and probability of words in a hyponymic relation. For this purpose, a number of diverse domain specific corpora have to be built, reflecting different interests relevant **for** particular semantic fields. Currently, we have tested the above claims for a small Dutch corpus (12,125 word tokens) on horses, which has been extracted from the Internet. It mainly contains documents on horse riding, horse breeding and horse sports. We extracted from the corpus all the nouns that have been used to refer to *horses*. This resulted in a set of 69 nouns. Of these 37 are included in the Dutch wordnet and 32 nouns represent terminology or productive compounds. The corpus does not include biological classes. For this, «her corpora have to be created. Among the words that are in the Dutch wordnet we find many circumstantials. Remarkably, there are also quite a few words that

For all these nouns we calculated the probability that it occurs in the domain specific corpus and the probability that it occurs in a large domain-neutral corpus (40 million token corpus included the Celex database). The probability is obtained by dividing the frequency by the number of words with the same part of speech in the corpus (Sanfilippo 1997). The results of this calculation are shown in Table 1, where we only listed the nouns included in the Dutch wordnet. The second column gives the frequency of the word in the domain corpus and the fourth column the frequency according to the Celex database. The third and fifth column give the probability. The final column gives the relative probability by dividing the domain probability by the general probability.

Noun	Domain Tokens	Probability in Domain	Celex Tokens	Probability accord. to Celex	Domain/ Celex
kind (child)	2	0.0009	40727	0.3281	0.007
ouder (parent)	2	0.0009	9079	0.0731	0.033
dier (animal)	4	0.0017	7772	0.0626	0.064
volwassene (adult)	1	0.0004	1499	0.0121	0.133
vader (father)	25	0.0107	24408	0.1966	0.106
zoon (son)	9	0.0038	7997	0.0644	0.125
ster (star)	3	0.0013	2604	0.0210	0.153
grootmoeder (grandmother)	5	0.0021	1214	0.0098	0.492
rivaal (rival)	1	0.0004	146	0.0012	1.351
huisdier (pet)	3	0.0013	320	0.0026	1.235
vos (fox)	3	0.0013	314	0.0025	1.258
woudloper (trapper)	1	0.0004	81	0.0007	2.410
stamvader (ancestor)	1	0.0004	76	0.0006	2.564
veulen (foal)	2	0.0009	145	0.0012	2.027
zuster (sister)	7	0.0030	423	0.0034	1.856
beestje (animal pet)	3	0.0013	129	0.0010	3.008
paard (horse)	169	0.0721	6675	0.0538	2.484
kampioen (champion)	5	0.0021	175	0.0014	3.315
halfbloed (half-bred)	1	0.0004	31	0.0002	6.061
ruin (gelding)	3	0.0013	59	0.0005	6.349
nakomeling (descendant)	11	0.0047	151	0.0012	7.362
merrie (merry)	20	0.0085	274	0.0022	7.119
trekpaard (draught horse)	2	0.0009	24	0.0002	11.111
rijpaard (riding horse)	15	0.0064	55	0.0004	22.535
dekhengst (stud)	7	0.0030	24	0.0002	25.000
volbloed (thoroughbred)	15	0.0064	40	0.0003	28.571
nestor	9	0.0038	19	0.0002	34.483
koetspaard (coach horse)	1	0.0004	2	0.0000	50.000
hengst (stallion)	137	0.0584	197	0.0016	41.194
springpaard (jumper)	12	0.0051	2	0.0000	86.667

Table 1: Domain-specific frequency and Domain-independent frequency for horse-referring nouns.

The table is sorted for the relative probability. The order clearly shows that specific circumstantials (including *rijpaard* (riding horse) and *trekpaard* (draught horse)) have a much stronger probability than the general categories: *paard* (horse) and *dier* (animal). This is due to the fact that the latter words have a high frequency in general. However, when we look at more abstract circumstantials, such as *kampioen* (champion), *rivaal* (rival), we see that they have a low probability too but only a slightly higher general frequency than the specific circumstantial. Typically, these words can be applied to diverse things, which explains the frequency in the general corpus (the specific circumstantials can only be used for horses). When we extend the data for other domains, we expect that the probability of general circumstantials will remain the same, whereas the probability for the specific *horse-dependent* circumstantials will be zero. A similar effect of domain-specific correlation is expected for expert terminology.

The most important conclusion is that the probability of general circumstantial is not significant for the domain. This means that their usage is not predicted by standard cluster techniques. A statistical search that only takes the frequency of word in documents into account will not match the general circumstantials with the query word *paard* (horse). A traditional hierarchy that does not differentiate between the different hyponymy types will not expand to these words either. Undifferentiated expansion to co-hyponym and co-hyperonyms will also include disjoint classes, and may lead to the selection of the whole hierarchy. A hierarchy as in Figure 4 will however allow activation or expansion to the non-exclusive hyperonyms too.

Conclusions

In this paper we described different ontological organizations, by making a distinction between ontologies of lexicalized words and expressions of languages and ontologies that include artificial levels of

classification. We explained the (dis)advantages of these ontologies and argued for the necessity to limit ourselves to lexicalized classes in a multilingual database as EuroWordNet.

Within the lexicalized classifications, we made a further distinction between conventional, circumstantial and specialized classifications. Different characteristics of these have been described, which can be accounted for by a different encoding in the wordnets. We showed that it is possible to develop a hybrid system in which the different classifications can be integrated but can be fine-tuned in terms of the probability or likelihood of classes to refer or apply to more specific meanings in certain contexts. This is important information not only for Information Retrieval applications but also for Machine-Translation tools and language-learning and generation tools.

The current proposal incorporates an explicit marking of different type of hyponymic dependencies. We also described a procedure for extending this explicit linking with a mutual domain-correlation score to indicate domain-consistency of hyponymy-relations. This mutual domain-correlation score can be extracted from co-occurrence frequencies in the same document for all word pairs that have a direct hyponymic relation in the hierarchy. Consistent pairs have parallel probability restrictions to the same documents (domains) and thus prioritize the classification scheme that they represent.

Acknowledgements

Bibliographical References

Agirre, E. and Rigau, G. (1996) *Word sense disambiguation using conceptual density.* Proceedings of COLING 1996.

Alonge, A., N. Calzolari, P. Vossen, L.Bloksma, I. Castellon, T. Marti, W. Peters (fc.) *The Linguistic Design of the EuroWordNet Database.* Computer and the Humanities, 1998.

Church, K. and P. Hanks (1990) *Word association norms, mutual information, and lexicography,* Computational Linguistics, 161: 22-29, 1990.

Cruse, D. A. (1986) *Lexical Semantics,* Cambridge, Cambridge University Press.

Dik, S. (1978) *Stepwise Lexical Decomposition,* Lisse, Peter de Ridder Press

Grefenstette, G. (1994) *Corpus-Derived First, Second and Third-Order Word Affinities.* Proceedings of Euralex 1994, 279-290.

Gruber, T.R. (1992) *Ontolingua: a Mechanism to Support Portable Ontologies.* Report KSL 91-66. Stanford University. 1992

Guarino, N. (1997) *Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration.* In M. T. Pazienza (ed.) Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer Verlag: 139-170. 1997.

Hovy, E. (fc.) *What would it Mean to Measure an Ontology?* ISI, University of Southern California.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross and K. Miller. (1990) *Five Papers on WordNet.* CSL Report 43. Cognitive Science Laboratory. Princeton University.

Lenat, D. and R. Guha (1990) *Building Large Knowledge-based Systems. Representation and Inference in the CYC Project.* Addison Wesley 1990

Peters, W., P. Vossen, P. Diez-Orzas, G.Adriaens (fc.) *Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index.* In Computer and the Humanities, 1998.

Pustejovsky J. (1995) *The Generative Lexicon.* The MIT Press. Cambridge, MA. 1995

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of IJCAI 1995.

Rodriquez, H., S. Climent, P. Vossen, L. Bloksma, A. Roventini, F. Bertagna, A. Alonge, W. Peters (fc.), *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology..* Computer and the Humanities, 1998.

Sanfilippo, A. 1997, *Using Semantic Similarity to Acquire Co-occurrence Restrictions from Corpora.* Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, 1997. 82-89.

Sowa, J. (fc.) *Knowledge Representation: Logical, Philosophical, and Computational Foundations,* PWS Publishing Co., Boston.

Vossen, P. (1995) Grammatical and Conceptual Individuation in the Lexicon, PhD. Thesis, University of Amsterdam, IFOTT, Amsterdam. 1995.

Vossen, P., P. Boersma, A. Bon, T. Donker (1995) *A flexible semantic database for information retrieval tasks.* Proceedings of the AI'95.

Vossen, P., P. Diez-Orzas, W. Peters *(1997) The Multilingual Design of EuroWordNet..* Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, 1997. 1-8.