EuroWordNet: building a multilingual database with wordnets for European languages.

Piek Vossen.

1. Introduction

All the knowledge and information in the Information Society is useless unless we are able to communicate with the keepers of it: computer systems. Most of the information they hold is stored as text and pictures, which people may understand but computers do not. It is clear that morpho-syntactic analysis and speech-processing will not bring us very far to exploit this information. Statistic techniques have been more successful, especially in Information Retrieval, mainly because they are computationally tractable, they do not rely on expensive resources and they can be applied to any domain that contains large quantities of text. Nevertheless, the benefits of shallow statistic processing are limited and the time seems ripe for exploring a more content-driven processing of information.

It is only fair to say that the area of semantics and interpretation includes many hurdles and pitfalls that make it difficult to define its limits and scope. Meaning is said to be fuzzy, complex, context-dependent, knowledge-dependent, and ambiguous. Still, some recent projects, such as the development of WordNet, EDR, MikroKosmos, Cyc, have shown that it is possible to develop large-scale resources involving part of the required knowledge that are feasible. These resources are being used, showing that it is not necessary to know the full scope of the problem to do useful things. Even stronger, we will only be able to tackle the full problem when we start dealing with parts of it in a realistic applied environment.

In Europe, these resources are not (yet) available in most languages. An additional problem is the multilinguality. The European Information Society not only needs these resources in every language but also a mapping across every language-resource. This is an absolute prerequisite for the successful development of the European Information Society. EuroWordNet directly addresses this problem by developing a multilingual database with wordnets for a large set of European languages. Each of these wordnets is structured along the same lines as the Princeton WordNet around the notion of a synset. A synset is a set of synonymous word meanings, between which basic semantic relations are expressed, such as hyponymy (car – vehicle), meronymy (wheeled vehicle – wheel), cause (kill – die). In addition to the relations between synsets, the so-called language-internal relations, each synset in EuroWordNet is also linked to some Inter-Lingual-Index or ILI, thus constituting a multi-lingual database (see Figure 1.). This ILI is an unstructured list of concepts, so-called ILI-records, mainly taken from WordNet1.5 but adapted to improve the matching of synsets across languages. Although the ILI as such will not be structured in terms of semantic relations between the concepts, it will nevertheless give access to a shared top-ontology and a domain-ontology. These ontologies are applied to particular sets of ILI-records, and, in principle, apply to any language-specific synset that is related to these ILI-records.

Architecture of the EuroWordNet Data Structure

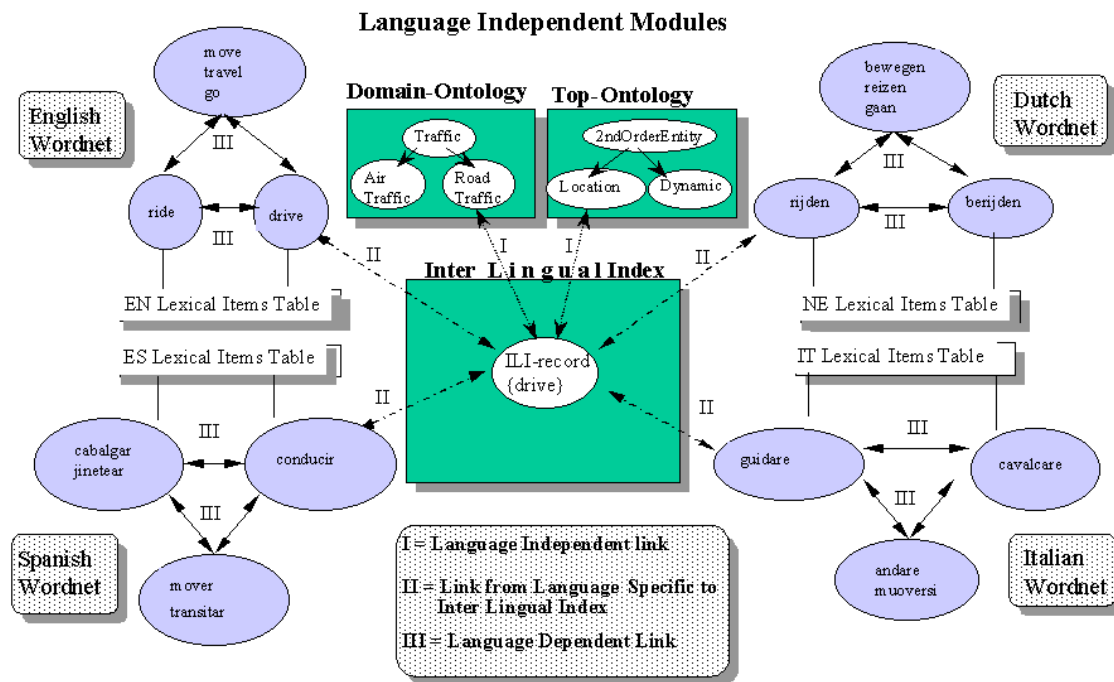**Language Independent Modules**

Figure 1: Overview of the EuroWordNet Database Design

Via the ILI it is possible to go from a synset in one wordnet to the synsets in the other wordnets that are related to the same ILI-record, and to compare the lexical semantic structures. A comparison of a large set of wordnets will give an indication of the differences in the relations across the wordnets. These differences can either be inconsistencies or they can point to language-specific differences of the resources. The fact that we link a whole series of wordnets to the ILI makes it possible to develop a more fundamental view on these differences, helping to understand how language-specific the wordnets are and pointing to areas where work remains to be done. The proportion of lexical semantic relations that is shared by a large number of wordnets gives a good indication about the quality of the relations. Special interfaces have been developed in the EuroWordNet database to carry out this kind of comparison.

The first consortium of the project (LE2-4003) has worked on the Dutch, Italian and Spanish wordnets, while the English wordnet was only adapted for relations which have not been covered in the Princeton WordNet1.5. Recently, the project has been extended (LE4-8328) to include French, German, Czech and Estonian. The wordnets are as much as possible built from existing resources, covering the general, generic vocabulary of the languages. The languages in the first project (LE2-4003) aim at a size of 30K synsets and 50K word senses. The languages in the extension will aim at a set of 15K synsets and 30K word meanings. Finally, the wordnets will be validated by 3 users in (cross-linguistic) Information Retrieval (IR) applications. The validation tools as such will not be developed, instead, the wordnets will be loaded in existing IR systems. Further information on the project and the participants can be found at the EuroWordNet WWW-site, http://www.let.uva.nl/~ewn.

2. Wordnets as autonomous language-specific networks

An important characteristics of the project is that the wordnets are treated as autonomous systems of language-internal relations. This will give us the flexibility to develop the wordnets relatively independently, which is needed because each group has a different starting point in terms of resources, tools and databases. However, there is also a more-fundamental reason why we take this position. Each wordnet represents a unique network of relations, due to the lexicalization patterns that are specific to the languages. For example, in the Dutch wordnet we see that *hond* (dog) is both classified as *huisdier* (pet) and *zoogdier* (mammal). However, there is no equivalent for *pet* in Italian, and likewise the Italian *cane*, which is linked to the same synset *dog*, is only classified as a *mammal* in the Italian wordnet. In EuroWordNet, we take the position that it must be possible to reflect such differences in lexical semantic relations. The wordnets are seen as linguistic ontologies rather than ontologies for making inferences only. In an inference-based ontology it may be the case that a particular level or structuring is required to achieve a better control or performance, or a more compact and coherent structure. For this purpose it may be necessary to introduce artificial levels for concepts which are not lexicalized in a language (e.g. *natural object*, *external body parts*), or it may be necessary to neglect levels which are lexicalized but not relevant for the purpose of the ontology. A linguistic ontology, on the other hand, exactly reflects the lexicalization and the relations between the words in a language. It is a "wordnet" in the true sense of the word and therefore captures valuable information about the expressiveness of languages: what is the available fund of words and expressions in a language.

The difference is illustrated in Figure 2, where the hyponymic structure of WordNet1.5 reflects a combination of lexicalized and non-lexicalized categories and the Dutch Wordnet only contains categories lexicalized in the language. In WordNet1.5 we see that the synset for *object* is first subdivided into two subclasses *artifact* and *natural object*, of which the latter is not a lexicalized expression in English (which you would expect in a dictionary) but rather a regularly composed expression. The class *artifact* has an important subclass *instrumentality*, which is used to group related synsets such as *implement*, *device*, *tool* and *instrument* below a common denominator. Such a grouping seems helpful to organize the hierarchy and predict the functionality of the subclasses. However, it does not give correct predictions about the substitutability of the nouns: you cannot refer to *containers*, *boxes*, *spoons*, and *bags* using the noun *instrumentality* in English.

Wordnet1.5 Ontology

object
- artifact, artefact (a man-made object)
  - block
  - instrumentality
    - implement
      - tool
    - container
      - box
      - spoon
      - bag
    - device
      - instrument
- natural object (an object occurring naturally)
  - body

Dutch Wordnet Ontology

voorwerp {*object*}
- blok {*block*}
- tas {*bag*}
- bak {*box*}
- lichaam {*body*}
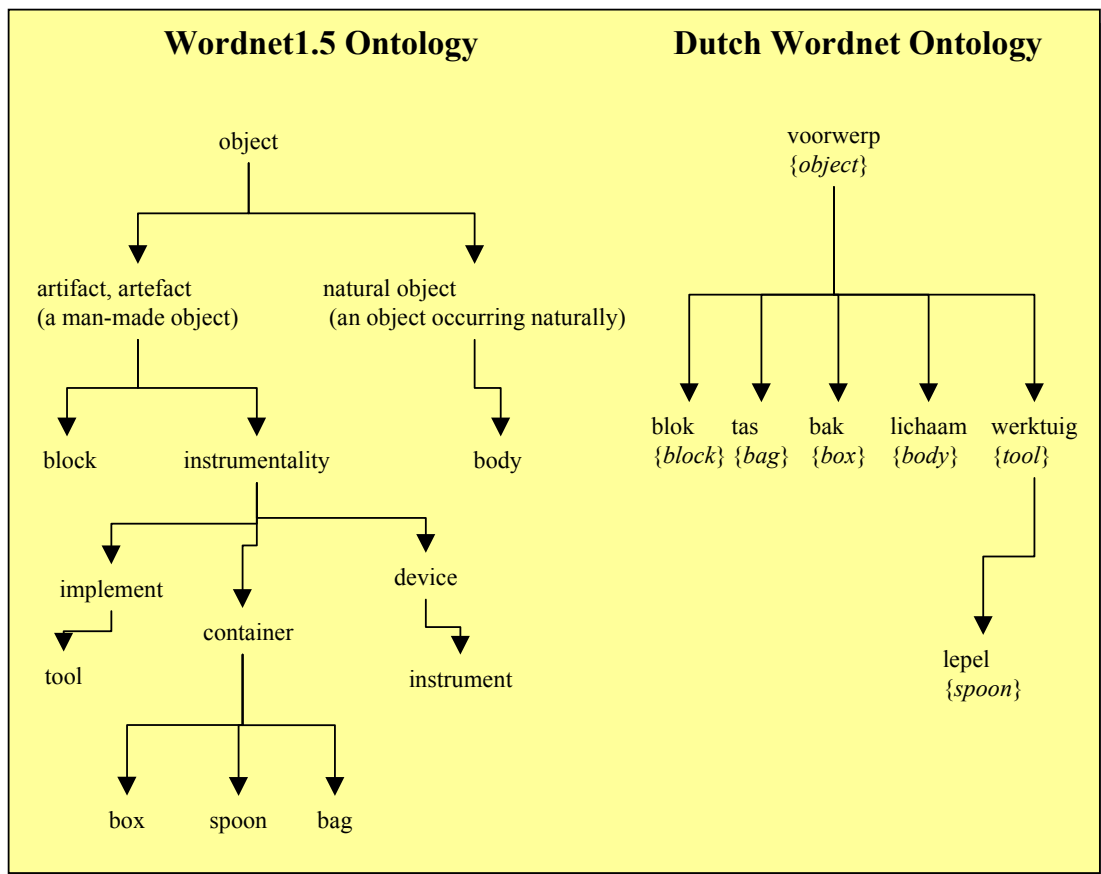- werktuig {*tool*}
  - lepel {*spoon*}

Figure 2: Lexicalized and Non-lexicalized levels in wordnets.

In the Dutch hierarchy, we see that artificial levels such as *natural object* and *instrumentality* have not been used. Furthermore, there are no exact equivalents for *artifact* and *container* in Dutch.[1] As a result of this, we get a much flatter hierarchy in which particular properties such as *natural*, *artificial* and *functionality* cannot be derived. On the other hand, the network correctly predicts the expressive capacity of Dutch because it only includes the legitimate words (and expressions) of the language. We could invent new classes and expressions in Dutch to capture different generalizations, we could even take over the WordNet1.5 classes, but there are no a priori criteria to decide what are useful classes and what are not. We may end up with adding any conceivable semantic property as a class to create very rich inheritance structures, or we may take over all possible classifications from all the other wordnets. However, this would destroy the wordnet as a network of legitimate expressions in a language and it would still not automatically give us a good conceptual ontology for inheriting properties. Besides that, it is possible to extend the database with a separate language-neutral ontology which takes care of the inferences and is well designed for that purpose. When this ontology is linked to the ILI, all the wordnets can access the classifications there to find the correct inferences for the synsets. The wordnets then provide the precise mapping of the language-specific vocabulary on this ontology. To get at such ontology, we are cooperating with the Ansii Group on Standardizing Ontologies, which is developing a standardized Reference Ontology.

---

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

3. The top-down building of the wordnets.

A drawback of the flexible design described above is that the interpretation and coverage of the wordnets may easily drift apart. There is no guarantee that we cover the same conceptual areas or that we encode the relations in the same way. To minimalize this danger, the wordnets are developed tow-down starting with a shared set of so-called Base Concepts. These Base Concepts have been selected for their importance in the local wordnets. Importance has been measured in terms of the number of relations and the position in the hierarchy. The more relations or the higher the position, the more important a meaning is. All meanings which play a major role in at least two wordnets have been selected. This has resulted in a set of 1059 Base Concepts, represented as WordNet1.5 synsets. The Base Concepts have been described using a top-ontology with 63 basic semantic distinctions (Top-Concepts) such as Substance, Object, Artifact, Natural, Function, Dynamic, Static, Cause, Location, Experience. The top-ontology has been based on other available ontologies and has been adapted to reflect the diversity of the Base Concept selection. The classification of the Base Concepts in terms of the Top-Ontology provides a common framework for the development of the individual wordnets by the different sites.

The actual building of the separate wordnets then takes place along the following steps:

1. The selection of a well-defined set of word meanings.
2. The encoding of lexical semantic relations and equivalence relations for this set.
3. Converting the data to the EuroWordNet import format.
4. Loading the data in the EuroWordNet database.
5. Comparing the wordnets for particular subsets.
6. Revising the wordnets in the EuroWordNet database.
7. Extending the first selection.

First, each group has determined the synsets that most closely represent the common Base Concepts in their local language, given the available resources. This selection has been extended with other meanings which are important in the local wordnets but which are not part of the common set of Base Concepts. This set of meanings has been classified in the local wordnets in terms of their hyperonyms, resulting in a unified tree. Note that these classifications may be different from wordnet to wordnet and still be compatible with the top-ontology classification. In addition to this top-layer, we have included those hyponyms that are also (important) hyperonyms of more specific meanings. Together this selection represents the core of each wordnet with the most important meanings on which the remainder of the vocabulary depends. To summarize, each core wordnet includes at least:

1. The best representatives for the 1059 Base Concepts.
2. Other meanings important for the local wordnet.
3. Hyperonyms for the local Base Concepts.
4. Most important hyponyms of the local Base Concepts.

The core wordnets are specified at least for synonymy, hyponymy and their equivalence relation to the ILI. Optionally, any other salient relation has been encoded to *interconnect* the meanings in the wordnet. Because of the importance for the total wordnets, the manual work has been focused on these cores. The extension from the core wordnets will by done in a top-down direction, using semi-automatic techniques. Currently, the top-ontology, the Base Concepts and the core-wordnets have been finalized for Dutch, Italian and Spanish. The data have been loaded in the EuroWordNet database and are being compared. From the

comparison in the EuroWordNet database it may follow that particular relations or word meanings are missing, that they have to be revised or that equivalence relations are not correct. This will lead to a modification of the core wordnets. In the remainder of the project, the cores will be extended and the other languages will be added. The new languages will first develop similar core wordnets and extend them in a later phase.

A separate task is the adaptation of the ILI. In practice it turns out to be difficult to find a precise matching between a synset in the local wordnet and a synset in the ILI (mostly synsets taken from WordNet1.5). Especially the Base Concepts, which often are vague and polysemous, are difficult to match. In many cases there will be a many-to-many matching or there will not be an equivalent concept in the ILI. To improve the matching, the ILI is adapted. There are two types of modification:

1. Adding of new concepts which are missing
2. Creating sense-groups between closely related senses or regular polysemy

The addition of new concepts is necessary to enable a precise mapping of synsets across wordnets in cases that there is no such concept in WordNet1.5. For example, if only the Spanish and Italian wordnet include a meaning for some type of *wine*, the new concepts should make it possible to specify the equivalence between Spanish and Italian despite the absence in English.

The sense-groups are necessary to deal with inconsistent and fuzzy sense-differentiation across the lexical resources. We often see that resources only specify one out of several meanings that can be distinguished (often on a regular basis): e.g. "embassy" as an *institute* or as a *building*. This may mean that concepts cannot be linked across languages because different meanings are represented: i.e. either the *institute* or the *building*. To relate these meanings across the wordnets, we extend the ILI with a globalized sense in which these meanings are grouped: "embasssy", both as a *building* and an *institute*. Each synset in the local wordnet that is linked to one of the more specific meanings will then get an additional equivalence link to the globalized meaning. These equivalence relations are differentiated from the normal equivalence relations so that it is possible to use these more global matches if a more precise matching gives no result.

4 Availability.

The shared components, such as the Top-Ontology, the ILI and the selection of Base Concepts, will be freely available. The same holds for the multi-lingual viewer that can be used to access the database version of these components and the wordnets. The language-specific wordnets will be property of the builders, in some cases in combination with the providers of the background resources. All components will be both available in database format and as ascii files. All wordnets can be licensed either from ELRA or from the owners. The core wordnets can be licensed as soon as the contracts are available from ELRA. We expect that the complete wordnets will become available beginning 1999. The tool for building and extending the databases can be licensed separately.