EuroWordNet: Linguistic Ontologies in a Multilingual Database.

Piek Vossen
Universiteit van Amsterdam
e-mail: Piek.Vossen@let.uva.nl
http://www.let.uva.nl/~ewn

Abstract

The aim of the EuroWordNet-project is to develop a database with wordnets for several European languages, similar to the Princeton WordNet1.5, which contains basic semantic relations between words in English. The wordnets will be linked to WordNet1.5 using equivalence relations constituting a multilingual database. In this paper we will give an overview of the basic principles and design of the database, and the overall procedure for building the wordnets.

Published in: Communication and Cognition for Artificial Intelligence. Special Issue "The creation and maintenance of electronic thesauri". 1998.

1. Introduction

The aim of EuroWordNet[1] is to develop a multilingual database with wordnets in several European languages. 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), etc., as is illustrated in Figure 1.

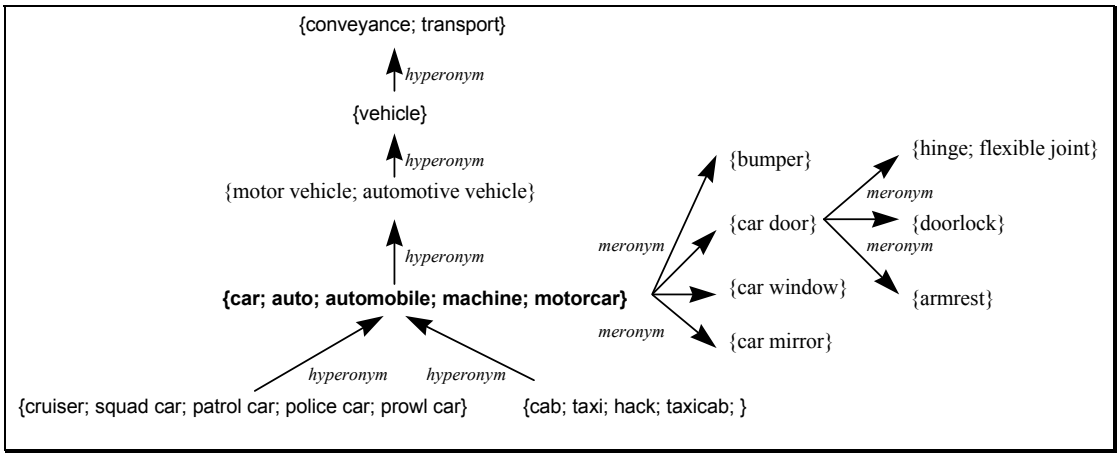

Figure 1: Meanings related to car#1 in WordNet1.5.

In this example, the set of synonyms {car; auto; automobile; machine; motorcar} is related to more general concepts (hyperonyms), such as {motor vehicle, automotive vehicle}, to more specific concepts (hyponyms), such as {cruiser; squad car; patrol car; police car; prowl car} and {cab; taxi; hack; taxicab}, to the parts it is composed of, such as {bumper}; {car door}, {car mirror} and {car window}. Each of these meanings or concepts is again related to other concepts as is illustrated for {motor

[1] EuroWordNet is funded by the EC as project LE2-4003 within the 4th Frame-Work of DG-XIII, Luxembourg. It is a joint enterprise of the University of Amsterdam (co-ordinator), the Fundacion Universidad Empresa (Madrid and Barcelona), Istituto di Linguistica Computazionale del CNR (Pisa), University of Sheffield, University of Tuebingen, University of Avignon, University of Tartu, University of Brno, Bertin (Paris), Memodata (Avignon), Rank Xeroc Research Center (Grenoble) and Novell Linguistic Development (Antwerp). Further information on the project can be found at: http://www.let.uva.nl/~ewn.

vehicle, automotive vehicle}, related to {vehicle}, and {car door} related to other parts: {hinge; flexible joint}, {armrest}, {doorlock}.

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 (what meanings can be seen as *vehicles*), 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 European wordnets will be stored in a central lexical database system and each meaning will be linked to the closest synset in the Princeton WordNet1.5, thus creating a multilingual database. In this database it will be possible to go from one meaning in a wordnet to a meaning in another wordnet, which is linked to the same WordNet1.5 concept. Such a multilingual database is useful for cross-language information retrieval, for transfer of information from one resource to another or for simply comparing the different wordnets. A comparison may tell us something about the consistency of the relations across wordnets, where differences may point to inconsistencies or to language-specific properties of the resources, or of the language itself. In this way, the database can also be seen as a powerful tool for studying lexical semantic resources and their language-specificity. The wordnets will as much as possible be built from available existing resources and databases with semantic information developed in various projects. This will not only be more cost-effective but will also make it possible to combine information from independently created resources, making the final database more consistent and reliable, while keeping the richness and diversity of the vocabularies of the different languages.

Currently, we are working on English, Dutch, Italian, and Spanish, and in the near future the database will be extended with German, French, Czech and Estonian. The size of the database will be around 30,000 comparable synsets in each language, corresponding with more or less 50,000 word meanings.[2] The vocabulary will comprise all the generic and basic words of the languages, that is: all word meanings needed to define more specific meanings and all words with high frequency in text-corpora. Some sub-vocabulary will be added for a domain to illustrate the possibility to integrate terminology in a general-purpose lexicon. The database will be tested in an existing multilingual information-retrieval application.

The structure of the database is first of all based on the so-called synset structure and relations of WordNet1.5 but contains some specific changes, mainly due to::

1) the multi-linguality of the database.
2) the nature of the information stored in the Machine Readable Dictionaries (MRDs) from which the EuroWordNet results will be derived and the possibility to (semi-)automatically extract this information.
3) to achieve maximal compatibility across the different resources.
4) to be able to maintain language-specific relations in the wordnets

In this paper we will discuss the basic principles and design of the database. In the next section we will first clarify our position with respect to the status of the semantic

[2] The languages added in the extension of the project (German, French, Czech and Estonian) will cover a smaller set of synsets.

relations that are expressed. As a multilingual database, it is unavoidable to take a clear position with respect to the language-specificity of semantic information. This position has consequences for the overall design of the database which is explained in section 3. Section 4 and 5 discuss the language-internal and equivalence relations, respectively. Finally, in section 6 we will explain the overall approach for building the wordnets, starting from a common framework. This framework is given by a top ontology applied to a set of shared Base Concepts, which play a major role in the wordnets.

2. Wordnets as Language-specific Ontologies

The most important difference of EuroWordNet with respect to WordNet is its multilinguality, which however also raises some fundamental questions with respect to the status of the monolingual information in the wordnets. In principle, multilinguality is achieved by adding an equivalence relation for each synset in a language to the closest synset in WordNet1.5. Synsets linked to the same WordNet1.5 synset are supposed to be equivalent or close in meaning and can then be compared. However, what should be done with differences across the wordnets? If 'equivalent' words are related in different ways in the different resources, we have to make a decision about the legitimateness of these differences. 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 conceptual ontologies. In a conceptual ontology it may be 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 therefore captures valuable information about the expressiveness of languages: what is the available fund of words and expressions in a language. Each wordnet should thus be seen as an autonomous language-specific structure. 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 *artefact* 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 *artefact* 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 organise the hierarchy and predict the functionality of the subclasses. However, it does not give correct predictions about the substitutionality of the nouns: you cannot refer to *containers*, *boxes*, *spoons*, and *bags* using the noun *instrumentality* in English.

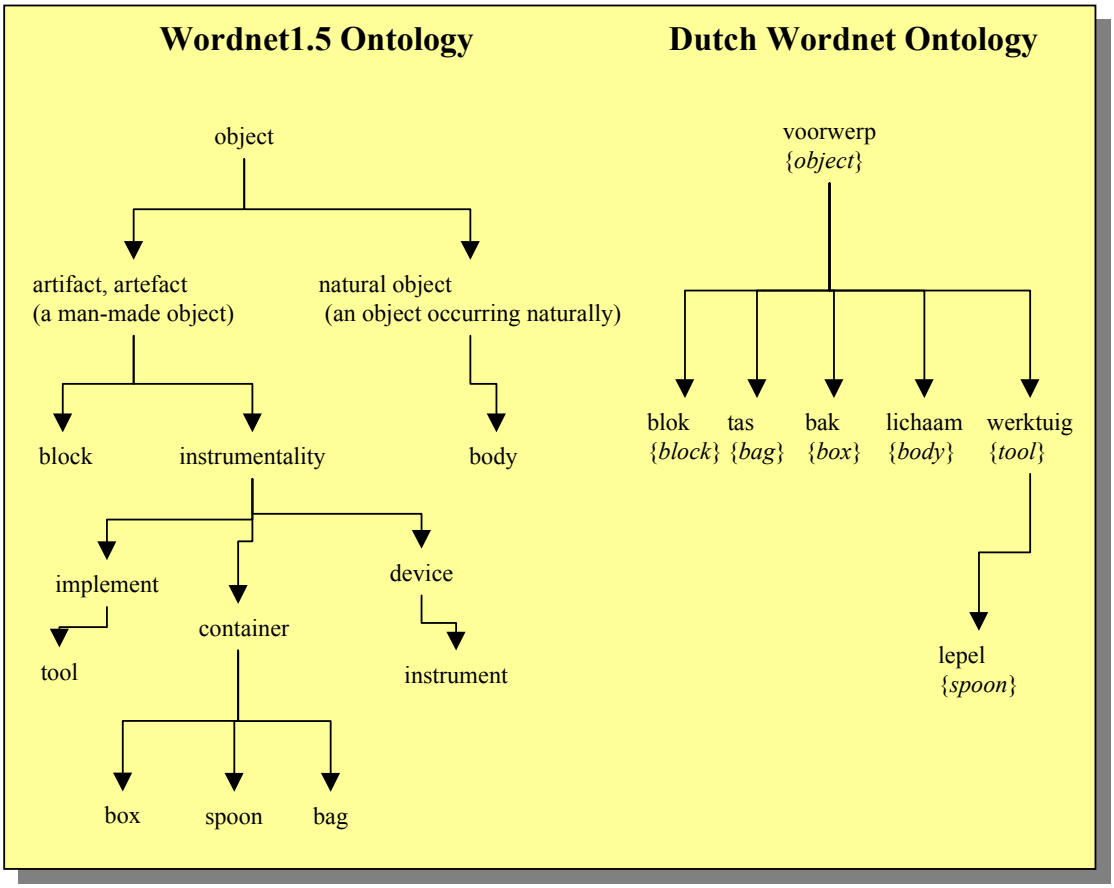

Figure 2: Lexicalized and Non-lexicalized levels in wordnets.

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 *artefact* and *container* in Dutch.[3] As a result of this we see 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 generalisations, 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. It is for example not always a good strategy to create many deep hierarchical levels, because the features cannot be applied outside the scope of the class. In the above example we see that *containers* are indirectly restricted to *artefacts*. It may very well be that natural objects or parts, such as *seedcase*, *lake*, *bladder*, *stomach*, have a similar function but they cannot be grouped below *container* because they belong to the disjoint class *natural object*.

Note that different hierarchical structures can still have the same effect in terms of inherited properties. In the case of the Dutch wordnet, it is possible to express the role as a *container* by a separate role-relation to the verb *bevatten* (to

[3] The word container is in Dutch only used for big containers on ships or for big garbage cans.

contain) for each of the objects that has such a function. Furthermore, we can avoid an explosion of levels by allowing multiple hyperonyms, e.g. to both *container* and *artefact* or *container* and *natural object* in WordNet1.5., creating a tangled hierarchy instead of a tree.

It is not trivial to decide in general what makes a good ontology [Gruber 1992], especially not from a multilingual perspective. However, it is possible to approximate the set of words and expressions in a language and to decide on the relations between these. For that we use several principles. First of all, for every relation we formulated a test-sentence (in each language) which can be used to verify the relation between two words. These substitution sentences function as diagnostic frames [Cruse 1986]:

i.	a.	It is a fiddle therefore it is a violin.
	b	It is a violin therefore it is a fiddle.
ii.	a.	It is a dog therefore it is an animal.
	b	*It is an animal therefore it is a dog.
iii.	a	to kill (/a murder) causes to die (/ death) to kill (/a murder) has to die (/ death) as a consequence
	b	*to die / death causes to kill *to die / death has to kill as a consequence

From i. it follows that *fiddle* and *violin* are synonymous, from ii. it follows that *dog* is a hyponym of *animal* and from iii. that a cause relation holds between *kill* and *die*. In addition to these substitution tests, we use a Principle of Economy [Dik 1978] to ensure that all lexicalized levels are included in the hierarchical relations:

> If a word W_1 (animal) is the hyperonym of W_2 (mammal) and W_2 is the hyperonym of W_3 (dog) then W_3 (dog) should not be linked to W_1 (animal) but to W_2 (mammal).

This principle applies to all transitive relations (hyponymy, meronymy, cause, subevent). Finally, a Principle of Compatibility ensures that different relations are consistently applied to similar concepts:

> If a word W_1 is related to W_2 via relation R_1, W_1 and W_2 cannot be related via relation R_n, where R_n is defined as a distinct relation from R_1.

These tests and principles are formulated in such a way that they arouse anomaly when applied to words that violate the constraint. It is thus possible to verify the relations as a native speaker of the language without having to bother too much about the inheritance effects of the created network.

Additionally, it is possible to still link the wordnets as language-specific lexicalizations of concepts to another resource, such as Wordnet1.5 or CYC [Lenat and Guha 1990], which then gives the conceptual inferences for concepts in the wordnets. The mapping of a lexicon to such a conceptual ontology will differ from language to language but the inferences remain the same.

In addition to the theoretical motivation there is also a practical motivation to consider the wordnets as autonomous networks. To be more cost-effective, they will be (as much as possible) derived from existing resources, databases and tools. This

gives each of the sites a different starting point for building their local wordnet. It is therefore important that we allow for a maximum of flexibility in producing the wordnets and structures.

3. The overall design of the EuroWordNet database

To be able to maintain the language-specific structures and to allow for the separate development of independent resources we make a distinction between the language-specific modules and a separate language-independent module in the multilingual database. Each language module represents an autonomous and unique language-specific system of language-internal relations between synsets. Equivalence relations between the synsets in different languages and WordNet1.5 will be made explicit in the so-called Inter-Lingual-Index (ILI). Each synset in the monolingual wordnets will have at least one equivalence relation with a record in this ILI. Language-specific synsets linked to the same ILI-record should thus be equivalent across the languages, as is illustrated in Figure 3 for the language-specific synsets linked to the ILI-record {drive}.

Figure 3 further gives a schematic presentation of the different modules and their inter-relations. In the middle, the language-external modules are given: the ILI, a Domain Ontology and a Top Concept Ontology. The ILI consists of a list of so-called ILI-records (ILIRs) which are related to word-meanings in the language-internal modules, (possibly) to one or more Top Concepts and (possibly) to domains. The language-internal modules then consist of a lexical-item-table indexed to a set of word-meanings, between which the language-internal relations are expressed.

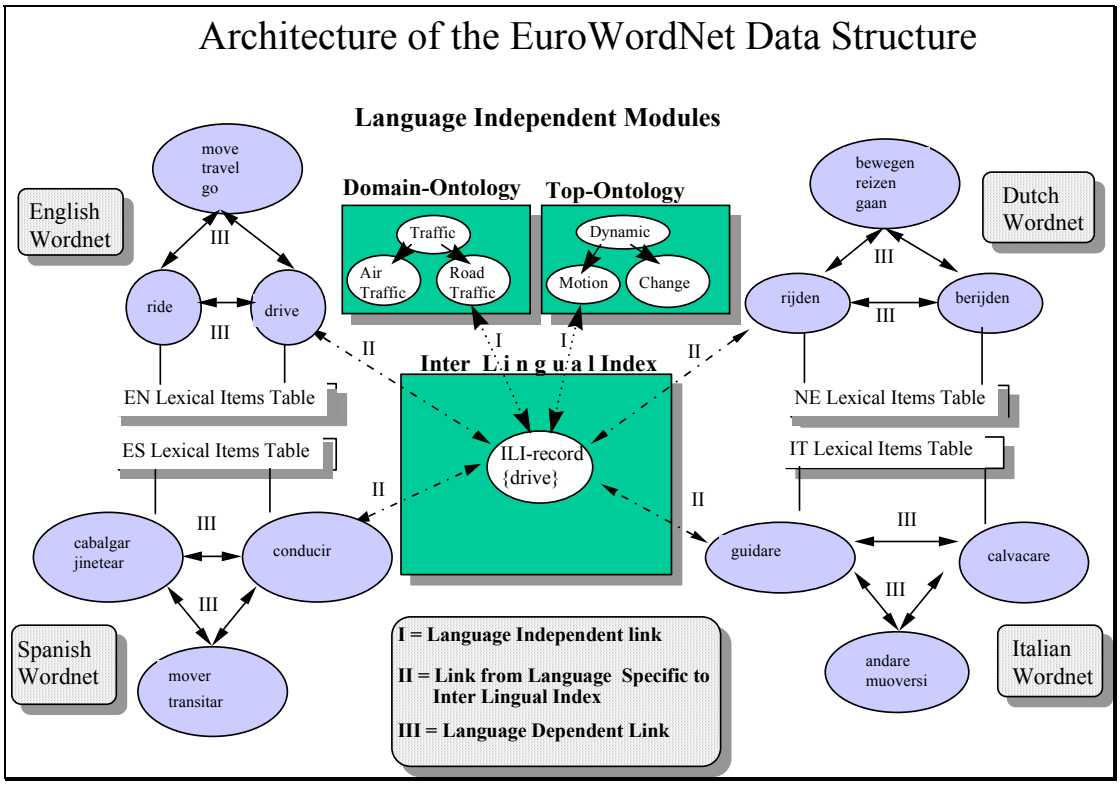

Figure 3. The global architecture of the EuroWordNet database.

The ILI is an unstructured list of meanings, where each ILI-record consists of a synset, a gloss specifying the meaning and a reference to its source. The ILI will initially consist of WordNet1.5 synsets but it will be adapted to provide a better matching with the language-specific synsets (see below section 5). No relations are maintained between the ILI-records as such. The development of a complete language-neutral ontology is considered to be too complex and time-consuming given the limitations of the project. The advantages of an unstructured ILI are:

- complex multilingual relations only have to be considered site by site and there will be no need to communicate about concepts and relations from a many-to-many perspective.
- future extensions of the database can take place without re-discussing the ILI structure. The ILI can then be seen as a fund of concepts which can be used in any way to establish a relation to the other wordnets.

It will nevertheless be possible to indirectly see a structuring of a set of ILI-records by viewing the language-internal relations of the language-specific concepts that are related to the set of ILI-records. Since WordNet1.5 is linked to the index in the same way as any of the other wordnets, it is still possible to recover the original internal organisation of the synsets in terms of the semantic relations in the WordNet1.5. A further discussion on the advantages and disadvantages of different multilingual designs and the ways of comparing the wordnets is given in [Vossen et al 1997a], and [Peters et al. fc.].

Some language-independent structuring of the ILI is nevertheless provided by two separate ontologies, which are linked to ILI records as is illustrated in Figure 3:

- the Top Concept ontology, which is a hierarchy of language-independent concepts, reflecting important semantic distinctions, e.g. Object and Substance, Process and Change;
- a hierarchy of domain labels, which are knowledge structures grouping meanings in terms of topics or scripts, e.g. Traffic, Road-Traffic, Air-Traffic, Sports, Hospital, Restaurant;

Both the Top Concepts and the domain labels can be transferred via the equivalence relations of the ILI-records to the language-specific meanings. In Figure 3, the Top Concept *Motion* is for example directly linked to the ILI-record *drive* and it therefore indirectly also applies to all language-specific concepts related to this ILI-record. Via the language-internal relations the Top Concept can be further inherited to all other related language-specific concepts. The main purpose of the ontologies is to provide a common framework for the most important concepts in all the wordnets. We will further discuss the top ontology in section 6.2 below.

4. The language-internal relations of EuroWordNet

The structure of the language-internal modules is based on the structure of WordNet 1.5:

- synonymous meanings are joined in a 'synset': e.g. *violin* and *fiddle*.
- language-internal relations are expressed between synsets.

The most important relations of WordNet1.5 are listed below with examples:

Relation	*POS-combination*	*Example*
ANTONYMY	adjective-to-adjective, verb-to-verb	Open/ close
HYPONYMY	noun-to-noun, verb-to-verb	Car/ vehicle, walk/ move
MERONYMY	noun-to-noun	Head/ nose
ENTAILMENT	verb-to-verb	buy/ pay
CAUSE	verb-to-verb	kill/ die

Table 1: Most important relations in WordNet1.5

Most of these relations are taken over with some changes and additions. The most important changes are:

- the use of labels on the relations
- explicit semantic relations across parts-of-speech
- a more global near-synonym relation
- sub-event relations instead of the use of entailment
- the interpretation of the cause-relation
- the use of role-relations between entities and events

These will be discussed in more detail below. For a more complete overview of the relations including the substitution tests in four languages see [Alonge 1996] and [Climent et al 1996].

4.1 Labels added to relations

Each language-internal relation may have one or more labels specify particular features or properties of the relation. The following labels are used:

- conjunction/disjunction
- non-factive
- reversed
- negation

The conjunction and disjunction label are used to explicitly mark the status of multiple relations of the same type occurring at a synset. In the Princeton WordNet1.5 the interpretation is not explicit. It is a matter of practice that e.g. multiple meronyms linked to the same synset are automatically taken as conjunctives: “all the parts together constitute the holonym *car* “. In the opposite case we see that parts, such as *door*, belonging to different kinds of holonyms are differentiated as different synsets or meanings of *door*:

door1 -- (a swinging or sliding barrier that will close the entrance to a room or building; "he knocked on the door"; "he slammed the door as he left") PART OF: doorway, door, entree, entry, portal, room access

door 6 -- (a swinging or sliding barrier that will close off access into a car; "she forgot to lock the doors of her car") PART OF: car, auto, automobile, machine, motorcar.

In more-traditional resources, similar relations are often expressed by explicit disjunction or conjunction of words in the same definition. In EuroWordNet disjunction and conjunction can therefore also explicitly be indicated by a label added to the relations:

{airplane}	*{door}*
HAS_MERO_PART: c1 {door}	HAS_HOLO_PART: d1 {car}
HAS_MERO_PART: c2d1 {jet engine}	HAS_HOLO_PART: d2 {room}
HAS_MERO_PART: c2d2 {propeller}	HAS_HOLO_PART:d3 {airplane}

Here c1, c2 and d1, d2, d3 represent conjunction and disjunction respectively, where the index keeps track of the scope of nested combinations. For example, in the case of *airplane* we see that either a *propeller* and *jet engine* constitute a part that is combined as the second constituent with *door*. Note that one direction of a relationship can have a conjunctive index, while the reverse can have a disjunctive one. Finally, when conjunction and disjunction labels are absent, multiple relations of the same types are interpreted as non-exclusive disjunction (and/or).

The label *Non-factive* is used to indicate that a causal-relation does not necessarily hold [Lyons 1977]:

- factive: event E_1 implies the causation of E_2, e.g.:
  "to kill causes to die":
  {kill} CAUSES {die}

- non-factive: E_1 probably or likely causes event E_2 or E_1 is intended to cause some event E_2:
  "to search may cause to find".
  {search} CAUSES {find} *non-factive*

Likewise, we can store different types of causal relations with different modal implications, and still differentiate the strength of the implication.

It is a requirement of the database that every relation has a reverse counter-part. However, there is a difference between relations which are explicitly coded as reverse relations and relations which are automatically reversed because of this requirement:

- if a *finger* is defined by reference to *hand* and *hand* is defined as a body part consisting of *fingers* then the relation is also conceptually bi-directional.
- if a *paper-clip* is made of *metal* then the reverse that *metal* can be shaped into a *paper-clip* but the latter is hardly relevant for explaining the meaning of *metal*.

To be able to distinguish between conceptually-dependent and automatically-reversed relations we therefore use the label *Reversed*:

{hand}	HAS_MERO_PART	{finger}	
{finger}	HAS_HOLO_PART	{hand}	
{paper-clip}	HAS_MERO_MADEOF	{metal}	
{metal}	HAS_HOLO_MADEOF	{paper-clip}	*reversed*

Obviously, expansion of words via a relation labelled *reversed* should be treated as less important than explicit relations which have no such label.

Finally, the negation label *Not* is used to explicitly express that a relation does not hold:

{monkey}	HAS_MERO_PART		{tail}
{ape}	HAS_MERO_PART	*not*	{tail}

This can be explicitly expressed using the negation label whereas we cannot express this as antonymy. Negation can be used to explicitly block certain key-word expansions. Below we will see more example of how the different labels can be used to differentiate properties of relations.

4.2 Explicit Cross-Part-Of-Speech relations

In Princeton WordNet nouns and verbs are not interrelated by basic semantic relations such as hyponymy and synonymy. The effect is that very similar synsets are totally unrelated only because they differ in part of speech (POS). This is illustrated by the following examples in which the noun *adornment* and the verb *adorn* have hyponymy-links which are not connected:

adornment 2	⇒	change of state -- (the act of changing something into something different in essential characteristics)
adorn 1	⇒	change, alter -- (cause to change; make different; cause a transformation; "The advent of the automobile may have altered the growth pattern of the city"; "The discussion has changed my thinking about the issue")

In the EuroWordNet project words of different parts of speech can be inter-linked with explicit xpos-synonymy, xpos-antonymy and xpos-hyponymy relations. The above examples will then explicitly be linked as follows:

{adorn V}	XPOS_NEAR_SYNONYM	{adornment N}

The advantages of such explicit cross-part-of-speech relations are:

- similar words with different parts of speech are grouped together.
- from an information retrieval point of view the same information can be coded in an NP or in a sentence. By unifying higher-order nouns and verbs in the same ontology it will be possible to match expressions with very different syntactic structures but comparable content (see the Sift project, LRE 62030, [Vossen and Bon 1996].
- by merging verbs and abstract nouns we can more easily link mismatches across languages that involve a part-of-speech shift. Dutch nouns such as “afsluiting”,

"gehuil" are translated with the English verbs "close" and "cry", respectively. In the combined hierarchy we can directly link "afsluiting" to "close" and "gehuil" to "cry".

As we will see below there are also more implicit relations across part of speech such as CAUSES, SUBEVENT and the ROLE relations.

4.3 Near_Synonym relation

In many cases there is a close relation between words but not sufficient to make them members of the same synset. Often this follows from the fact that the hyponyms linked to each of these words cannot be exchanged. This is shown in the following Dutch examples where near-synonyms of *instrument* have different classes of hyponyms. Typically, we see that *electrical devices* are linked to *apparaat* (apparatus) and *non-electrical devices* to *werktuig* (tool):

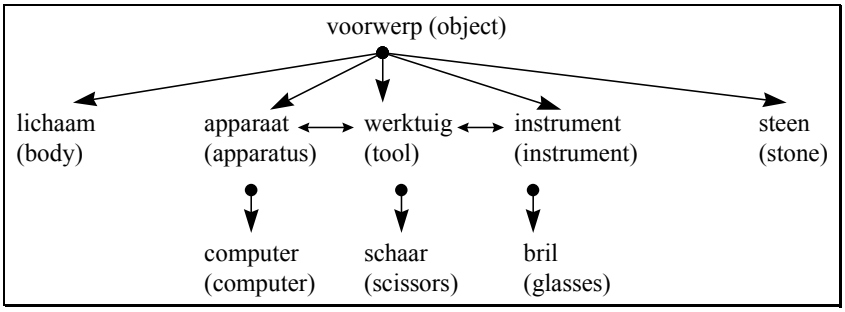

Figure 4: Near-synonyms in the Dutch wordnet.

If these words are joined in a single synset these groupings will be lost and we will get a counter-intuitive classification: i.e. *electric devices* are called *werktuig*, which native speakers will object to. However, if the meanings are kept separate it means that we cannot express the fact that they are much closer in meaning than other co-hyponyms such as *steen* (stone) and *lichaam* (body). For making the latter distinction the NEAR_SYNONYM relation can be used: hyponyms can be kept separate, but synsets can still be closely related, especially in contrast to other co-*hyponyms*. The distinction is relevant for Information Retrieval because it makes it possible to precisely predict which words can be expected to substitute other words in text (*apparaat* for *computer* but not *werktuig*), while it also enables to apply a more rough matching neglecting the distinction or transferring it to words in another language in which the distinction is not lexicalized.

4.4 SUBEVENT-relation

According to Fellbaum [Miller et al, 1990: 45] the entailment relation underlies all verbal relations: "the different relations that organise the verbs can be cast in terms of one overarching principle, lexical entailment". Next, lexical entailment is differentiated on the basis of the temporal relation between events and the direction of the implication or entailment:

a. + Temporal Inclusion (the two situations partially or totally overlap)

a.1 co-extensiveness (e. g., *to limp*/*to walk*) **hyponymy/troponymy**
a.2 proper inclusion (e.g., *to snore*/*to sleep*) **entailment**

b. - Temporal Exclusion (the two situations are temporally disjoint)
b.1 backward presupposition (e.g., *to succeed*/*to try*) **entailment**
b.2 **cause** (e.g., *to give*/*to have*)

In the actual database the relation Entailment is applied to those cases that cannot be expressed by the more specific hyponymy and cause relations. In that case at least the direction of the implication or entailment is indicated. In the case of *snore/sleep* the direction is from *snore* to *sleep*: i.e. *snore* implies *sleep* but not the other way around. In the case of *buy/pay* on the other hand *buy* implies *pay* but not the other way around.

In the EuroWordNet project, the differences in direction of the entailment can however be expressed by the labels factive and reversed. For example, 'backward presupposition' can be expressed by using the causal relation in conjunction with the factivity label:

{to succeed}	IS_CAUSED_BY	{to try}	*factive*
{to try}	CAUSES	{to succeed}	*non-factive*

The 'proper inclusion' can more adequately be described by means of the HAS_SUBEVENT/ IS_SUBEVENT_OF relation, where the implicational direction follows from the label *reversed*:

{to snore}	IS_SUBEVENT_OF	{to sleep}	
{to sleep}	HAS_SUBEVENT	{to snore}	*reversed*
{to buy}	HAS_SUBEVENT	{to pay}	
{to pay}	IS_SUBEVENT_OF	{to buy}	*reversed*

The SUBEVENT relation is very useful for many closely related verbs and appeals more directly to human-intuitions (parallel to part-whole relation of concrete entities).

4.5 The interpretation of the CAUSE relation

The causal relation is used in WN 1.5 when one verb refers to an event causing a resulting event, process or state referred to by the second verb (like in the case of *show*/*see*, *fell*/*fall*, *give*/*have*). The causal relation *only* holds between verbs and it should *only* apply to temporally disjoint situations [Miller et al, 1990: 54]. In the EuroWordNet database, on the other hand, the causal relation will also be applied across different parts of speech:

{to kill} V	CAUSES	{death} N	
{death} n	IS_CAUSED_BY	{to kill} v	*reversed*
{to kill } v	CAUSES	{dead} a	
{dead} a	IS_CAUSED_BY	{to kill} v	*reversed*
{murder} n	CAUSES	{death}n	
{death} a	IS_CAUSED_BY	{murder} n	*reversed*

In these examples we see that both verbs and *higher-order nouns* may denote events or processes (henceforth 'dynamic situations' or *dS*) which cause a resulting dynamic or non-dynamic situation which may again be referred to by a verb, higher-order noun, adjective or adverb. Furthermore, we distinguish three possible cases of temporal relationship between the (dynamic/non-dynamic) situations:

- a cause relation between two situations which are temporally disjoint: there is no time point when dS_1 takes place and also S_2 (which is caused by dS_1) and vice versa (e.g. *to shoot*/*to hit*);
- a cause relation between two situations which are temporally overlapping: there is at least one time point when both dS_1 and S_2 take place, and there is at least one time point when dS_1 takes place and S_2 (which is caused by dS_1) does not yet take place (e.g. *to teach*/*to learn*);
- a cause relation between two situations which are temporally co-extensive: whenever dS_1 takes place also S_2 (which is caused by dS_1) takes place and there is no time point when dS_1 takes place and S_2 does not take place, and vice versa (e.g. *to feed*/*to eat*).

These examples show that temporal disjointness is not a necessary criterion for allowing a causal relation. In practice, the Princeton database also contains causal relations between overlapping situations. Finally, the above example again show that close relations across part-of-speech are necessary to reflect the lexical variation within and across languages.

4.6 Role relations

In the case of many verbs and nouns the most salient relation is not the hyperonym but the relation between the event and the involved participants. These relations are expressed as follows:

{hammer}	ROLE_INSTRUMENT	{to hammer}	
{to hammer}	INVOLVED_INSTRUMENT	{hammer}	*reversed*
{school}	ROLE_LOCATION	{to teach}	
{to teach}	INVOLVED_LOCATION	{school}	*reversed*

These relations are typically used when other relations, mainly hyponymy, do not clarify the position of the concept network, but the word is still closely related to another word. Again, this results in close relations across part-of-speech (but possibly also between nouns when one of the nouns refers to an event, e.g. *tennis*). Obviously, the ROLE/INVOLVED relations represent a more distant connection than e.g. synonymy, hyponymy, and meronymy. However, using this relation we can group words in a different script-like or thematic way; compare the use of the domain-labels discussed above.

5. The Equivalence relations in EuroWordNet

In addition to the language-internal relations there are six different types of inter-lingual relations. The most straight forward relation is EQ_SYNONYM which applies to

meanings which are directly equivalent to some ILI-record. In addition there are relations for complex-equivalent relations, among which the most important are:

- EQ_NEAR_SYNONYM when a meaning matches multiple ILI-records simultaneously,
- HAS_EQ_HYPERONYM when a meaning is more specific than any available ILI-record.
- HAS_EQ_HYPONYM when a meaning can only be linked to more specific ILI-records.

The complex-equivalence relations are needed to help the relation assignment during the development process when there is a lexical gap in one language or when meanings do not exactly fit. The first situation occurs quite often, because the sense-differentiation in WordNet1.5 is much larger then in the traditional resources from which the other wordnets are being built. For example, in the Dutch resource there is only one sense for *schoonmaken* (to clean) which simultaneously matches with at least 4 senses of *clean* in WordNet1.5:

- {make clean by removing dirt, filth, or unwanted substances from}
- {remove unwanted substances from, such as feathers or pits, as of chickens or fruit}
- (remove in making clean; "Clean the spots off the rug")
- {remove unwanted substances from - (as in chemistry)}

The Dutch synset *schoonmaken* will thus be linked with an EQ_NEAR_SYNONYM relation to all these sense of *clean*.

The HAS_EQ_HYPERONYM is typically used for gaps in WordNet1.5 or in English. Such gaps can be genuine, cultural gaps for things not known in English culture, e.g. *citroenjenever*, which is a kind of gin made out of lemon skin, or they can be pragmatic, in the sense that the concept is known but is not expressed by a single lexicalized form in English. An example of the latter are Dutch *hoofd* which only refers to human head and Dutch *kop* which only refers to animal head, while English uses *head* for both. The HAS_EQ_HYPONYM is then used for the reversed situation, when wordnet1.5 only provides more narrow terms. In this case there can only be a pragmatic difference, not a genuine cultural gap. An example is Spanish *dedo* which can be used to refer to both *finger* and *toe*.

The ILI has to be the **superset** of all concepts occurring in the separate wordnets. The main reasons for this are:

- it should be possible to link equivalent non-English meanings (e.g. Italian *dito* and Spanish *dedo*) to the same ILI-record even when there is no English or WordNet equivalent.
- it should be possible to store glosses and domain-labels for non-English meanings, e.g.: all Spanish *bull-fighting* terms should be linked to ILI-records with the domain-label bull-fighting.

Initially, the ILI will only contain WordNet1.5 synsets but eventually it will therefore be updated with language-specific concepts, such as the gaps described above, using a specific update policy to avoid duplication of added concepts. In the case of the above

citroenjenever (lemon gin), we will thus add a new ILI-record, while the relevant language-specific synsets will get an additional EQ_SYNONYM relation to this record:

Dutch WordNet
citroenjenever
HAS_EQ_HYERPONYM {gin}
EQ_SYNONYM {citroenjenever, lemon gin: gin made of lemon skin}

From what has been said so far it follows that there can be a many-to-many mapping from local synsets to ILI-records. In the above cases, a single synset is linked to multiple target synsets in the ILI. This may either be with an EQ_NEAR_SYNONYM relation or with an HAS_EQ_HYPONYM/ HAS_EQ_HYPERONYM and with an EQ_SYNONYM to a new ILI-record. In the case of genuine, cultural gaps, the latter matching will probably also result in a situation where multiple synsets in a local wordnet are linked to the same ILI-record. Both the Dutch *citroenjenever* (lemon gin) and *jenever* (gin) are linked to the same general ILI-record *gin*, albeit with a HAS_EQ_HYPERONYM and EQ_SYNONYM relation respectively.[4] Furthermore, as explained above, it is possible to encode a NEAR_SYNONYM relation between synsets, which are close in meaning but cannot be substituted as easily as synset-members: e.g. *machine*, *apparatus*, *tool*. In that case it may very well happen that these near-synonyms are linked to the same target ILI-record, either with an EQ_SYNONYM or an EQ_NEAR_SYNONYM relation. Whenever a set of close meanings is related to multiple ILI-records we may very well find a situation where there is a many-to-many matching. Typically, we find such sets as *machine*, *apparatus*, *tool* in all the involved languages. We will then get a rather fuzzy matching from the wordnets to a global set of ILI-records.

Such a fuzzy matching is partially caused by differences in the sense-distinctions across resources. We already mentioned the example of *clean* where a single senses matches several senses in WordNet1.5. A danger resulting from this is that equivalent senses across wordnets are linked to close but different senses of the same word in WordNet1.5. A related phenomenon is that regular polysemy is inconsistently represented across resources. Although *university* may be used to refer to both the *institute* and the *building* we see that resources often only represent one of these meanings, or conflate them in a single meaning. This may again result in a situation that the local synsets for *university* cannot be matched across wordnets.

To limit this danger, we are extending the ILI with globalized senses that represent sets of more specific but related senses of the same word. In Figure 5, we see that the original linking of Dutch, Italian and Spanish equivalents for *university* has been extended with an HAS_EQ_METONYM relation to a new globalized ILI-record *university* which contains a reference to two more specific meanings. Via the HAS_EQ_METONYM relations the synsets can be retrieved despite of the different ways in which they are linked to the more specific synsets.

It is not necessary that the metonymy-relation also holds in the local language. In this example only the Dutch wordnet has two senses that parallel the metonymy-

[4] In general, we can state here that a combination of a simple and a complex equivalence relation to the same ILI-record should match the language-internal relation between local synsets.

relation in the ILI.[5] The Italian and Spanish example only list one sense (which may be correct or an omission in their resources). In the case of Spanish there are multiple equivalences to both senses of *university*, whereas the Italian synset is only linked to the *building* sense. The Spanish example is in fact equivalent to the new globalized ILI-record.

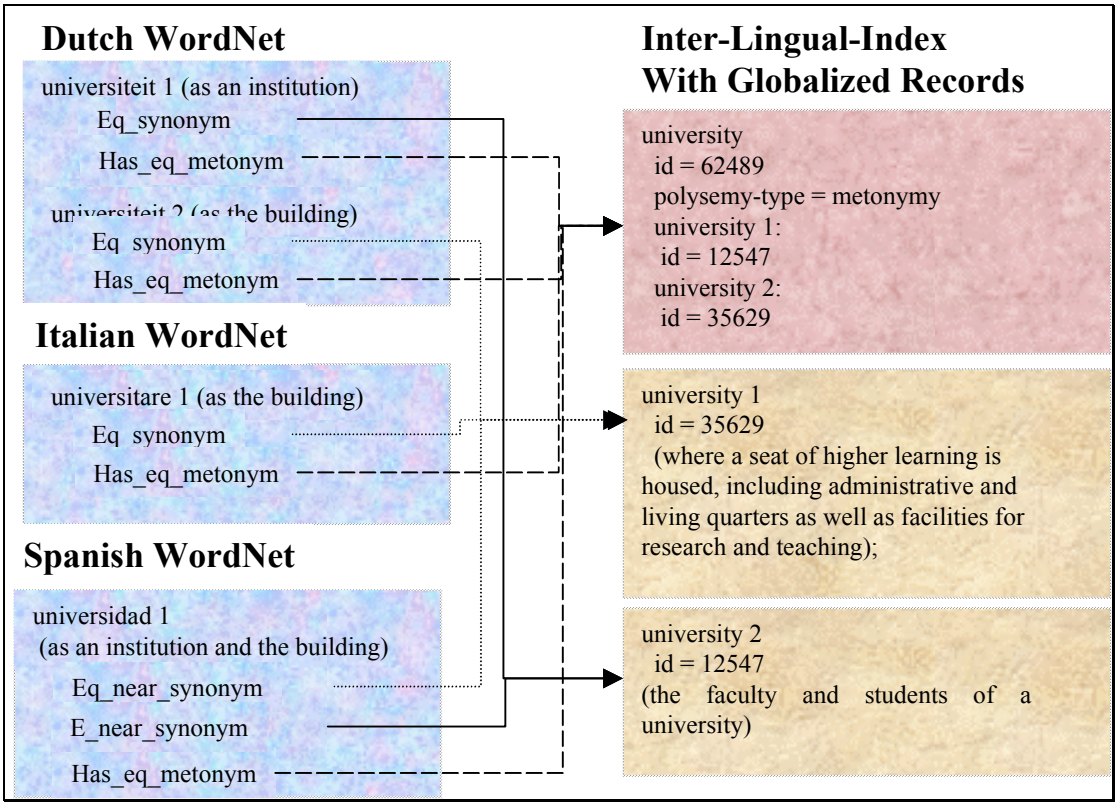

Figure 5: Inter-Lingual-Index with a globalized synset for *university*.

Similar globalized records are added for generalizations (*clean*) or verbal alternations such as causative and non-causative meanings: *he opens the door*, versus *the door opens* [Levin 1993]. In that case an HAS_EQ_GENERALIZATION or HAS_EQ_DIATHESIS relation will hold for synsets linked to more specific ILI-records that can be grouped in these ways. The generation of these equivalence relations is done fully automatically. After extending the ILI with more global concepts, the HAS-EQ_METONYM, HAS_EQ_GENERALIZATION or HAS_EQ_DIATHESIS will be automatically generated for all synsets which have at least one of the specific ILI-records in the globalized ILI-records as the target of an EQ_SYNONYM or EQ_NEAR_SYNONYM relation. There is no need for the local wordnet builders to consider each of these equivalence-extensions manually.

The global ILI-records will be derived using semi-automatic techniques from various data-sources. To some extent sense-groupings are already present in WordNet1.5, or they can be derived from the similarity of senses (in terms of the hyperonymic links, the glosses or the synset members). Furthermore, the fuzzy

[5] The relation between these two Dutch senses is now also expressed via the metonymy-equivalence relation to the more global ILI-record. The globalized ILI-record may also create metonymic relations between different forms which represent the same semantic relation, such as *universiteitsgebouw* (university building) in Dutch.

matching of local wordnets to WordNet1.5 also provides data for globalizing ILI-records.

6. The general approach for building the wordnets

In general, the wordnets are built in two major cycles as is illustrated in Figure 6 below. Each cycle consists of a building phase and a comparison phase:

1. Building a wordnet fragment
   1.1. Specification of an initial vocabulary
   1.2. Encoding of the language-internal relations
   1.3. Encoding of the equivalence relations
2. Comparing the wordnet fragments
   2.1. Loading of the wordnets in the EuroWordNet database
   2.2. Comparing and restructuring the fragments
   2.3. Measuring the overlap across the fragments

The building of a fragment is done using local tools and databases which are tailored to the specific nature and possibilities of the available resources. The available resources differ considerably in quality and explicitness of the data. Whereas some sites have the availability of partially structured networks between word senses, others start from genus words extracted from definitions that still have to be disambiguated in meaning.

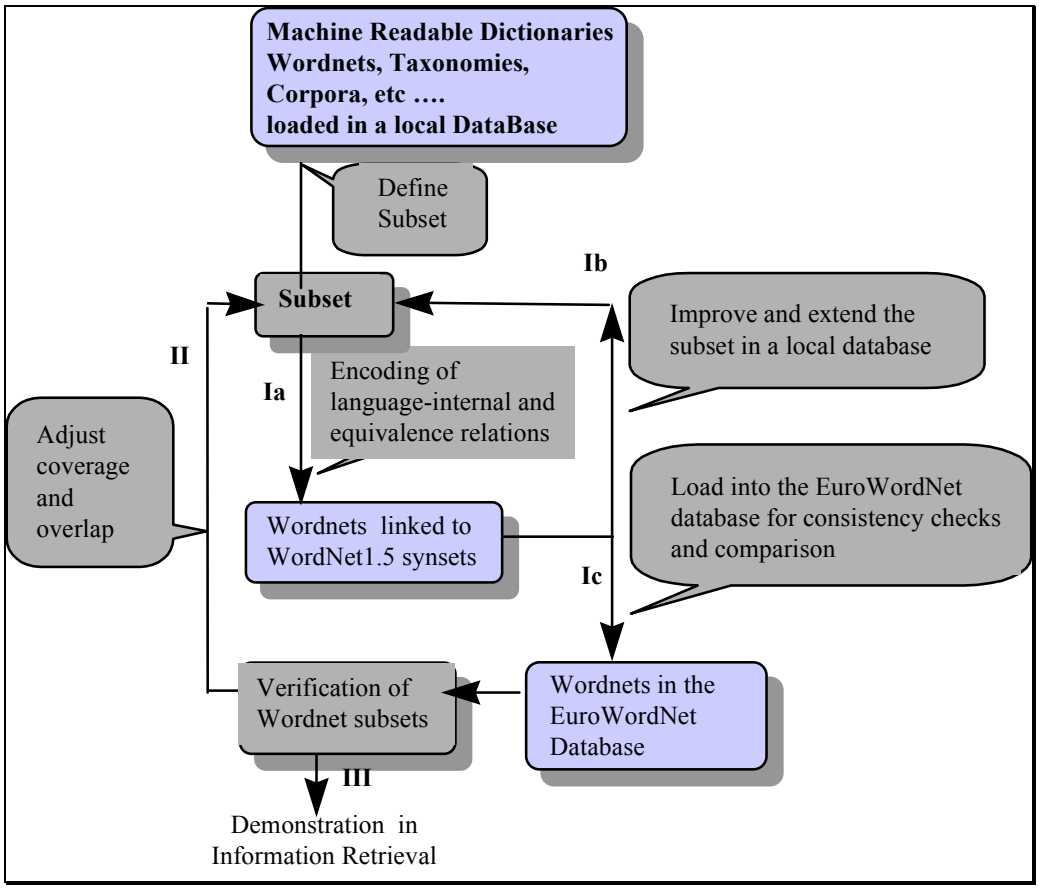

Figure 6: Overall procedure for building the wordnets.

After a production phase (step Ib and Ic in Figure 6) the results are converted to the EuroWordNet import format and loaded into the common database (step Ic). At that

point various consistency checks are carried out, both formally and conceptually. The database provides various ways of comparing the wordnets and finding structural mismatches. Synsets linked to the same Inter-Lingual-Index records should exhibit the same relations across the wordnets. The relations can directly be edited and the results can be exported. Special interfaces have been developed to compare and align the wordnets and trace down inconsistencies across the wordnets see [Cuypers and Adriaens 1997], [Vossen et al. 1997a] and [Peters et al. fc.], for further details.

At the end of each cycle there will be a verification phase by the users in the project. Feedback from the users can be incorporated in the next building cycle. At the end of the project the results will be used in a (cross-language) information retrieval application (phase III).

The development of the wordnets is done at separate sites in Europe, where each site as much as possible makes use of available resources. Since these resources vary across the sites, the groups have different starting points and likewise follow different strategies to develop their wordnets. Furthermore, the differences between the languages may also make it necessary to apply other methods. Dutch as a compounding language, exhibits different lexicalization patterns than Spanish and Italian, which also offers other possibilities for deriving relations.

It is therefore necessary to maintain a certain flexibility in the construction of the wordnets. However, flexibility should not lead to incompatibility of the results. Even though it must be possible that each wordnets represents a unique structure, it is not allowed that the interpretation of the relations and the overall conceptual coverage of the wordnets drifts apart. Only differences in lexicalization across the wordnets are acceptable.

Compatibility is to some extent guaranteed by loading the results in a common multilingual database, requiring a common format and syntax of the data. Nevertheless, to maximise the compatibility of the wordnets during the development we follow a top-down strategy where the different wordnets are developed starting from a shared set of so-called Base Concepts. These Base Concepts are selected on the basis of their role as fundamental building blocks in the different wordnets. The construction of the wordnets is then divided into two major phases:

1. the development of core wordnets by specifying the direct relations for the set of common Base Concepts.
2. The top-down extension of the core wordnets by adding more specific meaning that depend on the wordnet cores.

The development of the core-wordnets is mostly done manually, guaranteeing the highest quality and compatibility of the relations for the Base Concepts. Furthermore, by starting with the same set of Base Concepts we make sure that the same concepts are represented in all the wordnets, even though the lexicalizations around these cores may differ. The extensions are done using semi-automatic techniques, relying on the information stored in the available resources.

The interpretation of the relations is then further guided by the following measures:

(a) The availability of user-guides for building wordnets in each language:
    - the steps to encode the relations for a word meaning.
    - common tests and criteria for all the relations in the different languages.
    - overview of problems and solutions.

(b) Classification of the common Base Concept in terms of a Top-Ontology of 63 basic Semantic Distinctions

The guides for coding the relations are based on an extensive discussion of problematic meanings in the different wordnets. The tests for verifying the relations are formulated in the different languages. For a full specification of these tests and a discussion on the encoding of the relations see [Alonge 1996], [Climent et al. 1996], [Vossen et al, fc.]. In the next subsections we will shortly describe the selection of the Base Concepts and the Top Ontology in terms of which they have been classified, and we will give an overview of the tools and methods used for building the wordnets.

6.1 The selection of the Base Concepts

The Base Concepts play a major role in the different wordnets. This has been made operational by two selection criteria:

- The number of relations with other meanings
- The position in the hierarchy

The further procedure has globally been as follows:

(a) Each site determined the set of word meanings with most relations and high positions in the hierarchy.
(b) This set was extended with all meanings used to define the first selection.
(c) The local selections have been translated to WordNet1.5 equivalences, resulting in 4 lists of WordNet1.5 synsets (between 450 – 2000 synsets per selection).

These sets of WordNet1.5 translations have been compared. The intersection is extremely low: 30 synsets (24 nouns synsets, 6 verb synsets). Possible explanations for this are:

1. The individual selections are not representative enough.
2. There are major differences in the way meanings are classified, which have an effect on the frequency of the relations.
3. The translations of the selection to WordNet1.5 synsets are not reliable
4. The resources cover very different vocabularies

The last explanation is not very likely, since we only deal with general vocabulary. Differences in classifying (2) are of course legitimate. The quality of the translations (3) has been verified by looking at possible near-mismatches: cases where different senses of the same words have been selected. Closely related meanings (either selected or not) have been manually added to the intersection.

To further enlarge the coverage (1) we have taken the meanings which occur in at least two selections. This resulted in a set of 1024 synsets, divided as follows,

where 1stOrderEntities stand for concrete things, 2ndOrderEntities for properties and events and 3rdOrderEntities for ideas and propositions (see also the next section):

	Nouns	Verbs	Total
1stOrderEntities	491		491
2ndOrderEntities	272	228	500
3rdOrderEntities	33		33
Total	796	228	1024

Table 2: Total Set of shared Base Concepts

Each site has extended their local set of Base Concept so that these 1024 are as good as possible represented (e.g. about 50 Base Concepts could not be translated to equivalents in the Dutch wordnet, mostly technical classes such as *canine*, *vertebrate*). Each group started to encode the language-internal relations for this set of concepts, where each wordnet may exhibit different lexicalization patterns around them.

6.2 The EuroWordNet Top-Ontology

As suggested above the main purpose of the top ontology is to enforce the uniform encoding of the Base Concepts, by providing a common framework. Furthermore, using the top ontology we can divide the Base Concepts (BCs) into coherent clusters, which enables contrastive-analysis and discussion of closely related word meanings. Finally, the top ontology can be used to customize the database by assigning features to the top-concepts, irrespective of language-specific structures. In this sense it can also be used as an anchor point for connecting other ontologies to the Inter-Lingual-Index, such as e.g. CYC.

The first starting point for setting up the top ontology is that the wordnets are linguistic ontologies as discussed in section 2. We therefore used semantic distinctions which are common in linguistic paradigms: Aktionsart models [Vendler 1967, Verkuyl 1972, Verkuyl 1989, Pustejovsky 1991], entity-orders [Lyons 1977], Aristotle's Qualia-structure [Pustejovsky 1995]. Furthermore, we incorporated ontologies developed in previous EC-projects, which had a similar basis and are well-known in the project consortium: Acquilex (BRA 3030, 7315), Sift (LE-62030), [Vossen and Bon 1996].

The second starting point has been that the ontology should reflect the diversity of the set of common BCs, across the 4 languages. In this sense the classification of the common BCs in terms of the top-concepts should result in:

- homogeneous Base Concept Clusters
- average size of Base Concept Clusters

Homogeneity has been verified by checking the clustering of the BCs with their classification in WordNet1.5. The ontology has thus also been adapted to fit the top-levels of WordNet1.5. Furthermore, the clustering has been verified with the other language-specific wordnets. The criterion of cluster-size implies that we should not get extremely large or small clusters. In the former case the ontology should be further differentiated, in the latter case distinctions have to be removed and the BCs

have to be linked to a higher level. Finally, we can mention as important characteristics:

- the semantic distinctions should apply to both nouns, verbs and adjectives, because these parts-of-speech can be interrelated in the language-specific wordnets via a xpos_synonymy relation, and the ILI-records can be related to any part-of-speech.
- the Top Concepts are hierarchically ordered by means of a subsumption relation but there can only be one super-type linked to each Top Concept: multiple inheritance between top-concepts is not allowed.
- in addition to the subsumption relation, Top Concepts can have an opposition-relation to indicate that certain distinctions are disjoint, whereas others may overlap.
- there may be multiple relations from ILI-records to Top Concepts. This means that the BCs can be cross-classified in terms of multiple Top Concepts (as long as these are not disjoint): i.e. multiple inheritance from Top Concept to Base Concept is allowed.

It is important to realize that the Top Concepts (TCs) are more like semantic features than common conceptual classes. We typically find TCs for Living and for Part but we do not find a TC Bodypart, even though this may be more appealing to a non-expert. BCs representing *body parts* are now cross-classified by two feature-like TCs Living and Part. The reason for this is that the diversity of the BCs would require many cross-classifiying concepts where Living and Part are combined with many other TCs. These combined classes result in a much more complex system, which is not very flexible and difficult to maintain or adapt. Furthermore, it turned out that the BCs typically abstract from particular features but these abstractions do not show any redundancy: i.e. it is not the case that all things that are Living also always share other features.

An explanation for the diversity of the BCs is the way in which they have been selected. To be useful as a classifier or category for many concepts (one of the major criteria for selection) a concept must capture a particular generalization but abstract from (many) other properties. Likewise we find many classifying meanings which express only one or two TC-features but no others. In this respect the BCs typically abstract one or two levels from the cognitive Basic-Level as defined by [Rosch 1977]. So we more likely find BCs such as *furniture* and *vehicle* than *chair*, *table* and *car*.

The current ontology (version 1) is the result of 4 cycles of updating where each proposal has been verified by the different sites. The ontology now consists of 63 higher-level concepts, excluding the top. Following [Lyons 1977] we distinguish at the first level 3 types of entities:

1stOrderEntity
: Any concrete entity (publicly) perceivable by the senses and located at any point in time, in a three-dimensional space.

2ndOrderEntity
: Any Static Situation (property, relation) or Dynamic Situation, which cannot be grasped, heart, seen, felt as an independent physical thing. They can be

located in time and occur or take place rather than exist; e.g. continue, occur, apply

3rdOrderEntity

An unobservable proposition which exists independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten. E.g. idea, though, information, theory, plan.

According to Lyons, 1stOrderEntities are publicly observable individual persons, animals and more or less discrete physical objects and physical substances. They can be located at any point in time and in, what is at least psychologically, a three-dimensional space. The 2ndOrderEntities are events, processes, states-of-affairs or situations which can be located in time. Whereas 1stOrderEntities **exist** in time and space 2ndOrderEntities **occur** or **take place**, rather than exist. The 3rdOrderEntities are propositions, such as ideas, thoughts, theories, hypotheses, that exist outside space and time and which are unobservable. They function as objects of propositional attitudes, and they cannot be said to occur or be located either in space or time. Furthermore, they can be predicated as true or false rather than real, they can be asserted or denied, remembered or forgotten, they may be reasons but not causes.

The first division of the ontology is disjoint: BCs cannot be classified as combinations of these TCs. This distinction does however cut across the different parts of speech in that:

- 1stOrderEntities are always (concrete) nouns.
- 2ndOrderEntities can be nouns, verbs and adjectives, where adjectives are always non-dynamic (refer to states and situations not involving a change of state).
- 3rdOrderEntities are always (abstract) nouns.

All three parts-of-speech can be classified below the 2ndOrderEntity node. Note also that a BC may originally be a noun or verb in WordNet1.5 but may be associated with any part-of-speech in one of the local wordnets

Since the number of 3rdOrderEntities among the BCs was limited compared to the 1stOrder and 2ndOrder Entities we have not further subdivided them. The following BCs have been classified as 3rdOrderEntities:

theory; idea; structure; evidence; procedure; doctrine; policy; data point; content; plan of action; concept; plan; communication; knowledge base; cognitive content; know-how; category; information; abstract; info;

The 1stOrderEntities and 2ndOrderEntities are then further subdivided according to the following hierarchy:

Top	
1stOrderEntity	**2ndOrderEntity**
Origin Natural Living Plant Human Creature Animal Artifact **Form** Substance Solid Liquid Gas Object **Composition** Part Group **Function** Vehicle Symbol MoneySymbol LanguageSymbol ImageSymbol Software Place Occupation Instrument Garment Furniture Covering Container Comestible Building	**SituationType** Dynamic BoundedEvent UnboundedEvent Static Property Relation **SituationComponent** Cause Agentive Phenomenal Stimulating Communication Condition Existence Experience Location Manner Mental Modal Physical Possession Purpose Quantity Social Time Usage
3rdOrderEntity	

For a full description of the BCs and TCs see [Vossen et al. 1997b] and [Rodriquez et al. fc]. Here will only discus the most important distinctions.

6.2.1. Classification of 1st-Order-Entities

The 1stOrderEntities are distinguished in terms of four main ways of conceptualizing or classifying a concrete entity:

a) Origin: the way in which an entity has come about.
b) Form: as an a-morf substance or as an object with a fixed shape, hence the subdivisions Substance and Object.
c) Composition: as a group of self-contained wholes or as a part of such a whole, hence the subdivisions Part and Group.
d) Function: the typical activity or action that is associated with an entity.

These classes are comparable with Aristotle's Qualia roles as described in Pustejovsky's Generative lexicon, (the Agentive role, Formal role, Constitutional role and Telic Role respectively: [Pustejovsky 1995]) but are also based on our empirical findings to classify the BCs. BCs can be classified in terms of any combination of these four roles.

The main-classes are then further subdivided, where the subdivisions for Form and Composition are obvious given the above definition, except that Substance itself is further subdivided into Solid, Liquid and Gas. In the case of Function the subdivisions are based only on the frequency of BCs having such a function or role. In principle the number of roles is infinite but the above roles appear to occur more frequently in the set of common Base Concepts.

Finally, a more fine-grained subdivision has been made for Origin, first into Natural and Artifact. The category Natural covers both inanimate objects and substances, such as *stones*, *sand*, *water*, and all living things, among which *animals*, *plants* and *humans*. The latter are stored at a deeper level below Living. The intermediate level Living is necessary to create a separate cluster for natural objects and substances, which consist of Living material (e.g. *skin*, *cell*) but are not considered as *animate beings*. Non-living and Natural objects and substances, such as natural products like *milk*, *seeds*, *fruit*, are classified directly below Natural.

As suggested, each BC that is a 1stOrderEntity is classified in terms of these main classes. However, whereas the main-classes are intended for cross-classifications, most of the subdivisions are disjoint classes: a concept cannot be an Object and a Substance, or both Natural and Artifact. This means that within a main-class only one subdivision can be assigned. Consequently, each BC that is a 1stOrderEntity has at least one up to four classifications:

fruit:	Comestible (Function)	cell:	Part (Composition)
	Object (Form)		Living (Natural, Origin)
	Part (Composition)	life 1:	Group (Composition)
	Plant (Natural, Origin)		Living (Natural, Origin)
skin:	Covering (Covering)	reproductive structure 1	
	Solid (Form)		Living (Natural, Origin)
	Part (Constituency)		
	Living (Natural, Origin)		

Finally, with respect to Composition it needs to be said that only concepts that essentially depend on some other concept, are classified as either Part or Group. It is

not the case that all *persons* will be classified as Parts because they may be part of *group*. *Group*, on the other hand, typically depends on the elements as part of its meaning.

6.2.2. The classification of 2ndOrderEntities

As explained above, 2ndOrderEntities can be referred to using nouns and verbs (and also adjectives or adverbs) denoting static or dynamic Situations, such as *birth*, *live*, *life*, *love*, *die* and *death*. All 2ndOrderEntities are classified using two different classification schemes, which represent the first division below 2ndOrderEntity:

- the SituationType: the event-structure in terms of which a situation can be characterized as a conceptual unit over time;
- the SituationComponent: the most salient semantic component(s) that characterize(s) a situation;

The SituationType reflects the way in which a situation can be quantified and distributed over time, and the dynamicity that is involved. It thus represents a basic classification in terms of the event-structure (in the format tradition) and the predicate-inherent Aktionsart properties or nouns and verbs. The SituationComponents represent a more conceptual classification, resulting in intuitively coherent clusters of word meanings. The SituationComponents reflect the most salient semantic components that apply to our selection of Base Concepts. Examples of SituationComponents are: Location, Existence, Cause.

Typically, SituationType represents disjoint features that cannot be combined, whereas it is possible to assign any range or combination of SituationComponents to a word meaning. Each 2ndOrder meaning can thus be classified in terms of an obligatory but unique SituationType and any number of SituationComponents.

Following a traditional Aktionsart classification [Vendler 1967, Verkuyl 1972, Verkuyl 1989], SituationType is first subdivided into Static and Dynamic, depending on the dynamicity of the Situation:

Dynamic

Situations implying either a specific transition from one state to another (Bounded in time) or a continuous transition perceived as an ongoing temporally unbounded process; e.g. event, act, action, become, happen, take place, process, habit, change, activity. Opposed to Static.

Static

Situations (properties, relations and states) in which there is no transition from one eventuality or situation to another: non-dynamic; e.g. state, property, be. Opposed to Dynamic.

In general words, Static Situations do not involve any change, Dynamic Situations involve some specific change or a continuous changing.

Static Situations are further subdivided into Properties, such as *length*, *size,* which apply to single concrete entities or abstract situations, and Relations, such as *distance*, *space*, which only exist relative to and in between several entities (of the same order):

Property

Static Situation which applies to a single concrete entity or abstract Situation; e.g. colour, speed, age, length, size, shape, weight.

Relation

Static Situation which applies to a pair of concrete entities or abstract Situations, and which cannot exist by itself without either one of the involved entities; e.g. relation, kinship, distance, space.

Dynamic Situations are subdivided into events which express a specific transition and are bounded in time (BoundedEvent), and processes which are unbounded in time (UnboundedEvent) and do not imply a specific transition from one situation to another (although there can be many intermediate transitions):

BoundedEvent

Dynamic Situations in which a specific transition from one Situation to another is implied; Bounded in time and directed to a result; e.g. to do, to cause to change, to make, to create.

UnboundedEvent

Dynamic Situations occurring during a period of time and composed of a sequence of (micro-)changes of state, which are not perceived as relevant for characterizing the Situation as a whole; e.g. grow, change, move around, live, breath, activity, hobby, sport, education, work, performance, fight, love, caring, management.

We typically see that many verbs and nouns are under-classified for boundedness and sometimes even for dynamicity. This means that they can get a more specific interpretation in terms of a bounded change or an unbounded process when they are put in a particular context. A verb such as to walk names a bounded event when it is combined with a destination phrase, as in (a), but it is unbounded when it is combined with a location phrase as in (b):

a) He walked to the station (?for hours) (in 2 hours)
b) He walked in the park (for hours) (?in 2 hours)

The boundedness is made more explicit using duration phrases that imply the natural termination point of the change (*in 2 hours*) or explicitly do not (*for hours*).

The SituationComponents divide the Base-Concepts in conceptually coherent clusters. The set of distinctions is therefore based on the diversity of the set of common Base-Concepts that has been defined. As far as the set of Base Concepts is representative for the total wordnets, this set of SituationComponents is also representative for the whole. Note that adjectives and adverbs have not been classified in EuroWordNet yet. In this respect we may need a further elaboration of these components when these parts-of-speech are added. As said above, a verb or 2ndOrder noun may thus be composed of any combination of these components. However, it is obvious that some combinations make more sense than others.

The more specific a word is the more components it incorporates. Just as with the 1stOrderEntities we therefore typically see that the more frequent classifying nouns and verbs only incorporate a few of these components. In the set of common Base-Concept, such classifying words are more frequent, and words with many SituationComponents are therefore rare. Below are some examples of typical combinations of SituationComponents:

Experience + Stimulating + Dynamic+Condition (undifferentiated for Mental or Physical)

Verbs: *cause to feel unwell; cause pain*

Physical + Experience + SituationType (undifferentiated for Static/Dynamic)

Nouns: *sense; sensation; perception;*
Verbs: *look; feel; experience;*

Mental + (BoundedEvent) Dynamic + Agentive
Verbs: *identify; form an opinion of; form a resolution about; decide; choose; understand; call back; ascertain; bump into; affirm; admit defeat*
Nouns: *choice, selection*

Mental + Dynamic + Agentive
Verbs: *interpret; differentiate; devise; determine; cerebrate; analyze; arrange*
Nouns*: higher cognitive process; cerebration; categorization; basic cognitive process; argumentation; abstract thought*

Mental + Experience + SituationType (undifferentiated for Static/Dynamic)
Verbs: *consider; desire; believe; experience*
Nouns: *pleasance; motivation; humor; feeling; faith; emotion; disturbance; disposition; desire; attitude*

Relation+Physical+Location
Verbs: *go; be; stay in one place; adjoin*
Nouns: *path;course; aim; blank space; degree; direction; spatial relation; elbow room; course; direction; distance; spacing; spatial property; space*

Finally, it is important to realize that the Top Ontology does not necessarily correspond with the language-internal hierarchies. Each language-internal structure has a different mapping with the top-ontology via the ILI-records to which they are linked as equivalences. For example there are no words in Dutch that correspond with technical notions such as 1stOrderEntity, 2ndOrderEntity, 3rdOrderEntity, but also not with more down-to-earth concepts such as the Functional 1stOrder concept **Container**. These levels will thus not be present in the Dutch wordnet. From the Dutch hierarchy it will hence not be possible to simply extract all the *containers* because no Dutch word meaning is used to group or classify them. Nevertheless, the Dutch *containers* may still be found either via the equivalence relations with English *containers* which are stored below the sense of "container" or via the TopConcept clustering Container that is imposed on the Dutch hierarchy (or any other ontology that may be linked to the ILI).

6.3 The building process in EuroWordNet

Each of the groups starts with partially structured data, such as genus words extracted from dictionary definitions or taxonomies built from these genus relations. These structures are loaded into local databases. In the first cycle (see Figure 6 above) we encode the relations for the Base Concepts. The encoding proceeds from TC-cluster to TC-cluster, where we try to specify the language-internal relations around these clusters. The encoding of the Base Concepts is mostly done interactively using database editors that have specifically been designed for this purpose. In these editors, it is (among others) possible to browse through hierarchies, to encode relations between senses or groups of senses (supported by graphical interfaces), to export and import data, to implement heuristics for interpreting regular patterns, to view monolingual and bilingual dictionaries and to switch from resource to resource. For the interactive verification of the relations, the relation tests and the Top Ontology classifications are used.

In the second cycle, we extend the BCs top-down, where we try to enlarge the overlap across the wordnets and to include all the more frequent words from corpora (20,000 most frequent words). This work is done (semi)-automatically. At more specific levels, information is more clearly encoded in traditional dictionaries, and likewise can be extracted more easily using automatic techniques

In general, two global methods are followed for encoding the semantic relations (either manually or automatically):

> **Merge model**: the selection is done in a local resource and the synsets and their language-internal relations are first developed separately, after which the equivalence relations are generated to WordNet1.5.
>
> **Expand model**: the selection is done in WordNet1.5 and the WordNet.1.5 synsets are translated (using bilingual dictionaries) into equivalent synsets in the other language. The wordnet relations are taken over and where necessary adapted to EuroWordNet. Possibly, monolingual resources are used to verify the wordnet relations imposed on non-English synsets.

The Merge model results in a wordnet which is independent of WordNet1.5, possibly maintaining the language-specific properties. The Expand model will result in a wordnet which is very close to WordNet1.5 but which will also be biased by it. Whatever approach is followed also depends on the quality of the available resources.

If we take the Merge Model as starting point we, the following steps and modules (where some steps may already have been covered in other projects) are applied:

1. Parsing of definitions
2. Extraction of the syntactic head and other patterns from the parsed definitions.
3. Confirmation of the extracted relations
4. Disambiguation of extracted words
5. Mapping of monolingual entries to entries in bilingual resources
6. Mapping of translations to specific WordNet synsets

The parsing of the definitions is done using adapted general purpose parsers, specific definition parsers of more global pattern matching (see [Vossen et al 1989], [Vossen 1990], [Ageno et al. 1991a], [Ageno et al. 1991b], [Alonge 1991], [Calzolari et al. 1993], [Hagman 1992], [Marinai et al. 1993], [Montemagni 1992]). In all these cases, the structural analysis has to be converted to a conceptual relation, such as hyperonym, meronym or cause (see [Vossen 1991], [Hagman 1991], [Ageno et al. 1991b], [Vossen and Copestake 1993], [Oestling 1992], [Vossen et al 1995]. In most cases, the syntactic head represents the genus word, and other relations are extracted from specific patterns, like the following:

> X = Y which contains/ consists of/ has Z
> X HAS_MERONYM Z
> X = group of Z
> X HAS_MERO_MEMBER Z
> X = part of Z
> X HAS_HOLONYM Z
> X = to cause to Z

X CAUSES Z
X = to Y while Z
X IS_SUBEVENT_OF Z

These patterns are extracted using heuristics, many relations still have to be confirmed manually. The next step is to establish the exact sense of the word related in this way. This again can be done using heuristics that look for word overlap between the defined word (X) and the target word or manually using the above mentioned editors.

A specific step is building the synsets. In many cases the lexical resources (traditional dictionaries) already have explicit specifications of synonyms and antonyms. The main methods are:

- Single-word definitions and circular definitions are potential synonym, e.g.:
  rod = stick = pole = rod.
- Directly translating synsets from WordNet1.5 to synsets in another language (using bilingual dictionaries).
- Words with the same translations in bilingual dictionaries.

Synonymy is not always a clear notion and all heuristics need manual confirmation as well. Furthermore, synonyms typically occur at more general levels of the hierarchy, where the information is less reliable and a careful manual processing is required. At the more specific levels there will hardly be any synonyms.

The mapping of synsets in the local languages to WordNet1.5 equivalences is done in two steps. The first step consists of extracting the correct equivalences from bilingual dictionaries. For this purpose the senses of the monolingual resources have to be matched with the senses of the bilingual entries. Once the correct sense has been selected (where we again use heuristics for measuring the overlap in information in the monolingual and bilingual resources) the status of the translation has to be evaluated. In many cases the translation is a phrase or compound which does not occur in WordNet1.5. In that case, the local synset is linked to a more general synset in WordNet1.5 with a HAS_EQ_HYPERONYM relation (see [Ageno et al. 1993], [Vossen 1993] for details).

The next step is to find the relevant sense of the translation in the bilingual dictionary. For this some conceptual distance measurement is performed for each of the senses of the translation. Using this method, different senses of multiple translations are compared to select those senses with the closest distance in WordNet1.5. The distance-measurement is based on different parameters, such as the number of descendants, the depth of the hierarchy, the number of levels see [Agirre and Rigau 1996] for details. Alternatively, a translation can be compared with the translation of the related synsets in the source wordnet to find the best matching synset. If hyponyms and hyperonyms of a word are already translated, the translation with the closest conceptual distance to these translations is preferred. Again, this matching works reasonably well for the middle and bottom levels of the vocabulary, but is less reliable for very polysemous and vague words at the higher levels.

The result of the above steps is a monolingual wordnet fragment in one of the languages, where each synset has at least one equivalence relation to a WordNet1.5 synset. Each groups loads their fragment in the EuroWordNet database, where the equivalence relations to Wordnet1.5 are converted to equivalence relations with the

Inter-Lingual-Index. Successfully loaded wordnet fragments are then exchanged across the groups, where each group can compare the local wordnet with the other wordnet fragments. Such a comparison may result in a restructuring of the local wordnets, which can directly be done in the database. It may also result in adaptations of the Inter-Lingual-Index (ILI) to provide a better matching across the resources. These adaptations are carried out by one site. After such a comparison phase the resulting wordnets are exported and distributed with the adapted ILI to all partners. This represents the multilingual database at that phase in the project.

7. Conclusion

In this paper we have described the basic design features of the EuroWordNet database. We started by making a distinction between ontologies where classifications are based on the lexicalized units of languages and ontologies which contain non-lexicalized classes. The ontologies in EuroWordNet are wordnets in the true sense of the word, reflecting the unique lexicalization patterns of languages. This is also reflected in the multilingual design of the database where the wordnets are autonomous systems of language-internal relations. Multilinguality is then achieved by linking each synset to a concept in the so-called Inter-Lingual-Index: an unstructured fund of concepts, which represents the superset of all concepts occurring in the different wordnets. This Index gives access to all language-independent information: glosses, domain labels and semantic features. Such flexibility is needed because of the different starting points for the builders of the wordnets. We further described the language internal relations and the different equivalence relations used to encode the wordnets.

The modular design of the database makes it possible to develop the wordnets relatively independently. The last section explained how we nevertheless try to achieve compatible results by, among others, developing the wordnets top-down starting with a shared set of Base Concepts. These Base Concepts play a major role in the wordnets and are classified using a Top Ontology of lexical semantic distinctions. This ontology provides a common framework for encoding the language-specific configurations of the set of Base Concepts. The relations for the Base Concepts and their direct context are encoded manually in the local databases. In this way, the cores of the different wordnets are maximally compatible. Semi-automatic procedures are then used to extend the core-wordnets to more specific levels. The information for more specific meanings is more clearly encoded and can therefore more easily be extracted automatically.

References

[Ageno et al. 1991a] Ageno A. - I. Castellon - M.A. Marti - F. Ribas - G. Rigau - H. Rodriguez - M. Taule - M.F. Verdejo 1991a "The extraction of Semantic Information from MRDs", Esprit BRA-3030 Acquilex Working Paper 027/028

[Ageno et al. 1991b] Ageno A. - I. Castellon - M.A. Marti - F. Ribas - G. Rigau - H. Rodriguez - M. Taule - M.F. Verdejo 1991b "SEISD: User Manual. Guide to the Extraction and Conversion of Taxonomies", Esprit BRA-3030 Acquilex WP 030.

[Ageno et al. 1993] A. Ageno, F. Ribas, G. Rigau, H. Rodriguez and F. Verdejo 1993 TGE: Tlinks Generation Environment, Esprit BRA-7315 Acquilex2 Working Paper 8.

[Agirre and Rigau 96] Agirre E. and Rigau G. Word Sense Disambiguation using Conceptual Density, in proceedings of the 16th International Conference on Computational Linguistics (COLING'96). Copenhagen, Denmark. 1996.

[Alonge 1991] Alonge, A. 1991 "Extraction of information on Aktionsart from Verb Definitions in machine-readable dictionaries", in: Proceedings of the Avignon Conference on Natural Language Processing and its applications, 1991, Avignon.

[Alonge 1996], Alonge, A. (ed) Definition of the links and subsets for verbs, EuroWordNet Project LE4003, Deliverable D006. University of Amsterdam, Amsterdam. Http: //www.let.uva.nl/~ewn. 1996

[Bloksma et al. 1996] Bloksma, L., P. Díez-Orzas, and P. Vossen,The User-Requirements and Functional Specification of the EuroWordNet-project, EuroWordNet deliverable D001, LE2-4003, University of Amsterdam, Amsterdam. Http: //www.let.uva.nl/~ewn. 1996

[Calzolari et al. 1993] Calzolari, N., P. Cotoneschi, S. Montemagni and A. Spanu 1993 Extraction and Normalization of Noun Taxonomical Chains to Create a Thesaurical Set for Sense Disambiguation Tasks - Esprit BRA-7315 Acquilex2 Working Paper 15.

[Climent et al 1996] Climent, Salvador, Horacio Rodríguez, Julio Gonzalo (eds) Definition of the links and subsets for nouns of the EuroWordNet projec, EuroWordNet Project LE4003, Deliverable D005. University of Amsterdam, Amsterdam. Http: //www.let.uva.nl/~ewn. 1996

[Cuypers and Adriaens 1997]Cuypers, I. And G. Adriaens, Periscope: the EWN Viewer, EuroWordNet Project LE4003, Deliverable D008d012. University of Amsterdam, Amsterdam. Http: //www.let.uva.nl/~ewn. 1997

[Díez-Orzas et al 1996] [Díez Orzas, P. , Louw M. and Forrest, Ph, High level design of the EuroWordNet Database. EuroWordNet Project LE2-4003, Deliverable D007. 1996

[Dowty 1979] Dowty, D.R. Word meaning and Montague grammar. Dordrecht: Reidel.1979.

[Fellbaum 1990] Fellbaum, C. “English Verbs as a Semantic Net”, in: International Journal of Lexicography, Vol 3, No.4 (winter 1990), 278-301. 1990.

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

[Hagman 1991] Hagman, J. 1991 “Common and Odd Relations in Italian Dictionaries and their Treatment in Taxonomy Building”, Esprit BRA-3030 Acquilex Working Paper 044.

[Hagman 1992] Hagman, J. 1992 "Semantic parsing of Italian dictionary definitions", Esprit BRA-3030 Acquilex Working Paper 047.

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

[Levin 1993] Levin, B. (1993) *English Verb Classes and Alternations*. University of Chicago Press. Chicago.

[Lyons 1977] Lyons J. “Semantics”, Cambridge University Press, Cambridge. 1977

[Marinai et al. 1991] Marinai E. - C. Peters - E. Picchi 1991 "A Prototype System for the Semi-automatic Sense Linking and Merging of Mono- and Bilingual LDBs", paper presented at ACH/ALLC 91, Tempe, Usa, March 1991 and to be published in N. Ide and S. Hockey (eds.) Research in Humanities Computing. OUP.

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

[Montemagni 1992] Montemagni S. 1992 "Tailoring a broad coverage grammar for the analysis of dictionary definitions", in: H. Tommola & K. Varantola, T. Salmi- Tolonen & J. Schopp (eds.) Proceedings of the 5th Euralex International Congress on Lexicography. Tampere, Finland, 1992: 265-277.

[Oestling 1992] Oestling A. 1992 “Parts and wholes in dictionary definitions”, Esprit BRA-3030 Acquilex Working Paper 046.

[Peters et al. fc.] Peters, W., P. Vossen, P. Diez-Orzas, G. Adriaens, The Multilingual Design of the EuroWordNet database. In: Computer and the Humanities, Special Issue on EuroWordNet.

[Pustejovsky 1991] Pustejovsky, J. The syntax of event structure, Cognition, 41, 47-81. 1991.

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

[Rodriquez et al. fc.] Rodriguez, H., S. Climent, P. Vossen, L. Bloksma; A. Roventini, F. Bertagna, A. Alonge, W. Peters, The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In: Computer and the Humanities, Special Issue on EuroWordNet.

[Rosch 1977] Rosch, E. (1977) *Human Categorisation.* In N. Warren (Ed.) Studies in Cross-Cultural Psychology, Vol. I, pp. 1-49. Academic Press. London.

[Roventini and Antemi 1990] Roventini A. and D. Antelmi 1992 “Semantic Relationships within a Set of Verbal Entries in the Italian Lexical Database”, in: Euralex 1990, Proceedings, Bibliograf,

Barcelona.
[Vendler 1967]. Vendler, Z. Linguistics and philosophy. Ithaca: Cornell University Press. 1967.
[Verkuyl 1972] Verkuyl, H. On the compositional nature of the aspects. Dordrecht: Reidel. 1972.
[Verkuyl 1989] Verkuyl , H. “Aspectual classes and aspectual distinctions”, Linguistics and Philosiphy, 12, 39-94. 1989
[Vossen 1990] Vossen P. 1990 A parser-grammar for the meaning descriptions of the Longman dictionary of contemporary English. Technical Report NWO, project 300-169-007, University of Amsterdam.
[Vossen 1991] Vossen P. 1991 "Converting data from a lexical database to a knowledge base", Esprit BRA-3030 Acquilex Working Paper 027, University of Amsterdam.
[Vossen 1993] Vossen P. 1993 “Extracting equivalence relations for a multilingual lexical knowledge base”, Esprit BRA- 7315 Acquilex2Working Paper 014, University of Amsterdam.
[Vossen and Bon 1996] Vossen P. and A. Bon , “Building a semantic hierarchy for the Sift project”, Sift deliverable D20a, LRE 62030. Computer Centrum Letteren, University of Amsterdam. 1996.
[Vossen and Copestake 1993] Vossen P. - A. Copestake 1993 "Untangling definition structure into knowledge representation", in: E.J. Briscoe, A. Copestake and V. de Paiva (eds.) Default inheritance in unification based approaches to the lexicon. Cambridge: Cambridge University Press.
[Vossen et al 1989] Vossen P. - W.J. Meijs - M. den Broeder 1989 "Meaning and structure in dictionary definitions" in: B. Boguraev and T. Briscoe (eds.) Computational lexicography for natural language processing. London/New York: Longman: 171-190.
[Vossen et al 1995] Vossen P., P. Boersma, A. Bon, T. Donker 1995 “A flexible semantic database for information retrieval tasks”, in: Proceedings of the AI’95, June 26-30, Montpellier.
[Vossen et al. 1997a] Vossen, P., P. Diez-Orzas, W. Peters, The Multilingual Design of EuroWordNet. Published in: P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.
[Vossen et al. fc.] Vossen, P., L. Bloksma, C. Peters, A. Alonge, A. Roventini, E. Marinai, I. Castellon, T. Marti, G. Rigau, Encoding Language Internal and Equivalence Relations in EuroWordNet. In: Computer and the Humanities, Special Issue on EuroWordNet.
[Vossen et al.1997b] Vossen, P., H. Rodriguez, H., S. Climent, L. Bloksma, A. Roventini, F. Bertagna, A. Alonge, W. Peters, The Base Concepts and Top Ontology in EuroWordnet. EuroWordnet deliverable D017, LE2-4003. University of Amsterdam.