EuroWordNet: a multilingual database for information retrieval

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

Abstract

The aim of the EuroWordNet-project is the development of a database with wordnets for English, Spanish, Dutch and Italian, similar to the Princeton WordNet1.5, which contains basic semantic relations between words in English. The Dutch, Italian and Spanish wordnets will be linked to the WordNet1.5 using equivalence relations. The resulting multilingual database can directly be used in (multi-lingual) information retrieval. In this paper the basic characteristics of this database are described as far as relevant to this purpose.

Paper presented at the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.

1. Introduction

EuroWordNet[1] aims to build a multilingual database consisting of wordnets in several European languages (English, Dutch, Italian, and Spanish). Each language-specific wordnet is structured along the same lines as WordNet (Miller et al, 1990), i.e. synonyms are grouped in synsets, which in their turn are related by means of basic semantic relations. In addition each meaning will be linked with an equivalence relation to a WordNet1.5 synset, thus creating a multilingual database. The size of the database will be around 25,000 comparable synsets in each language, corresponding with more or less 50,000 word meanings. The vocabulary will comprise all the generic and basic words of the languages. For one or two domains sub-vocabulary will be added 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 EuroWordNet database itself 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. For that purpose the language-specific wordnets will be stored as independent language-internal systems in a central lexical database while the equivalent word meanings across the languages will be linked to each other via the equivalence relation with WordNet1.5.

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

1) the multi-linguality of the database.
2) the use of the database for Information Retrieval purposes.
3) 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.
4) to achieve maximal compatibility across the different resources.
5) to be able to maintain language-specific relations in the wordnets
6) the possibility for users to customise the database for their specific application without having to speak all the languages.

In this paper we will focus on the specific features of the database relevant to multi-lingual information retrieval. In the next section the multilingual design is discussed. In section 3 the language-internal relations are described as far as they are different from WordNet1.5.

---

[1] EuroWordNet is a 3-year project 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 Linguistica del Computazionale del CNR (Pisa), University of Sheffield, and Novell Linguistic Development (Antwerp). Further information on the project can be found at: http://www.let.uva.nl/~ewn.

2. The multilingual design of the EuroWordNet database

The major difference with WordNet1.5 is the multilingual design. All language specific wordnets will be stored in a central lexical database system. 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 synset or record in the ILI. Language-specific synsets linked to the same ILI-record should thus be equivalent across the languages.

The ILI starts off as an unstructured list or fund of WordNet1.5 synsets, and will grow when new concepts will be added which are not present in WordNet1.5. The advantages of the 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.

The actual internal organisation of the synsets by means of semantic relations can still be found in the WordNet database which is linked to it. The only organisation that will be provided to the ILI is via two separate ontologies which are linked to ILI records:

- the top-concept ontology: which is hierarchy of language-independent concepts, reflecting explicit opposition relations (e.g. Animate and Inanimate) and can be seen as representing the different semantic fields in the vocabulary of EuroWordNet.
- a hierarchy of domains labels which relate concepts on the basis of scripts or topics, e.g. «sports», «water sports», «winter sports», «military», «hospital».

Both the top-concepts and the domain labels can be transferred via the equivalence relations of the ILI-records to the language-specific meanings and, next, via the language-internal relations to any other meaning in the wordnets, as is illustrated in Figure 1 for the TCs *Object* and *Substance*.

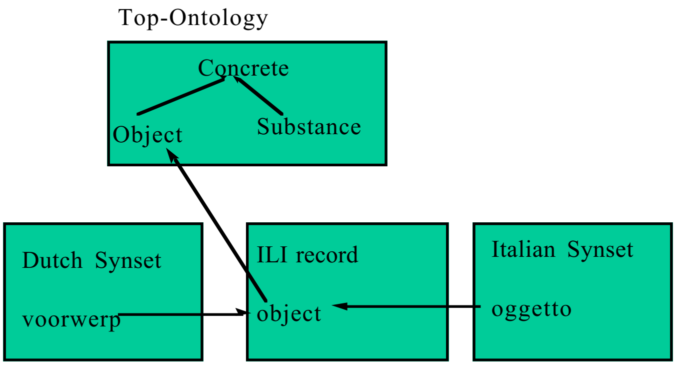

Figure 1.

Both hierarchies will enable a user to customise the database with semantic features without having to access the language-internal relations of each wordnet. Furthermore, the domain-labels can directly be used in information retrieval (also in language-learning tools and dictionary publishing) to group concepts in a different way, based on scripts rather than classification. Domains can also be used to separate the generic from the domain-specific vocabularies. This is important to control the ambiguity problem in Natural Language Processing.

The overall modular structure of the EuroWordNet database can be summed up as follows: first, there are the language modules containing the conceptual lexicons of each language involved. Secondly, there is the Language Independent Module which comprises the ILI, a Domain Ontology and a Top-Concept Ontology. Figure 2 gives a schematic presentation of the different modules and their inter-relations. In the middle, the language-external modules are given, made up by the list of so-called ILI-records (ILIRs) which are related to word-meanings (M) in the language-internal modules, (possibly) to one or more Top-Concepts (TCs) and (possibly) to domains (D).

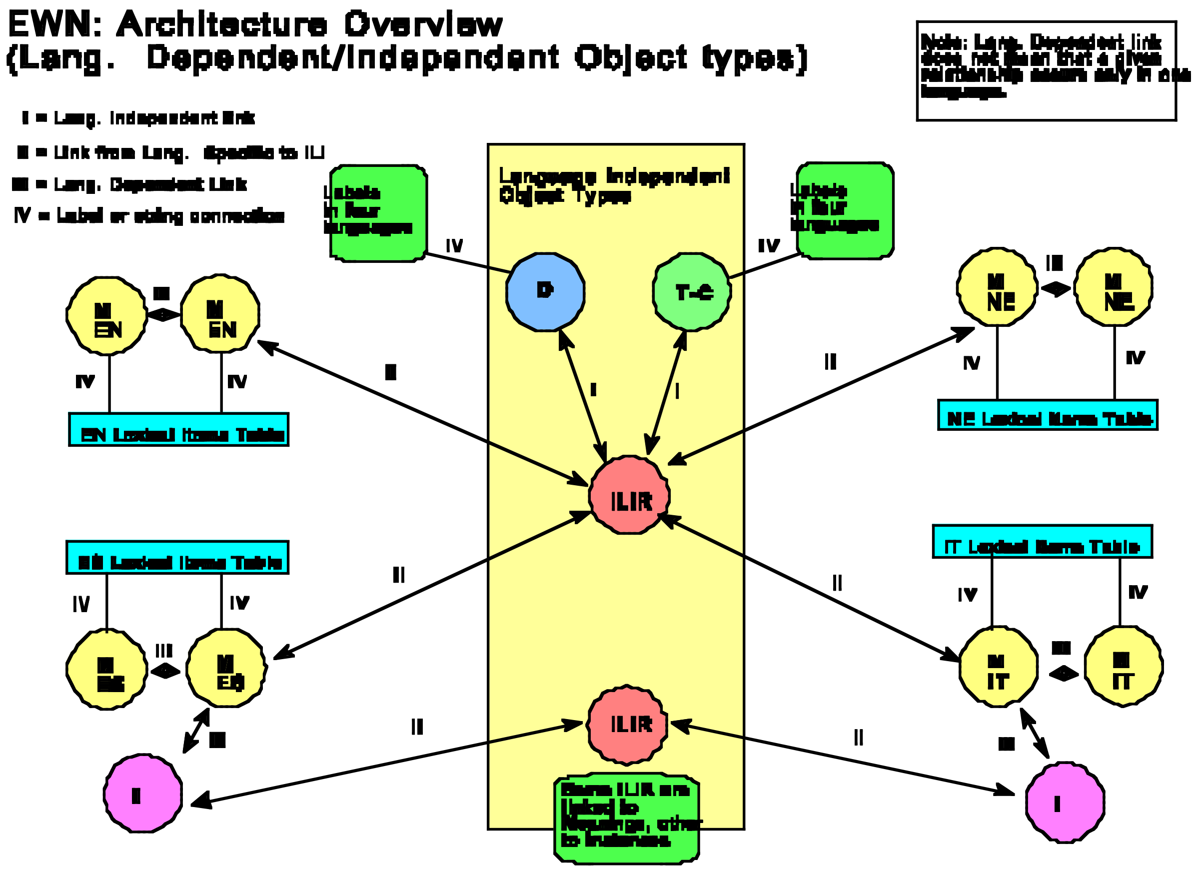

Figure 2. (Picture taken from Díez et al 1996)

The language-internal modules then consist of a lexical items table indexed to a set of word-meanings, between which the language-internal relations are expressed. Instances (I) occurring in a particular domain can be stored as language-specific concepts as well. This to enable the storage of language-specific names.

Figure 3 then illustrates how differences in lexicalizations are inter-linked via the ILIRs. Here we see that Italian *dito* and Spanish *dedo* can refer to both *fingers* and *toes*. These words are more general than the English meanings and are therefore linked to these ILIRs with a so-called eq_hyponym relation. The Dutch *hoofd* (human head) and *kop* (animal head) are on the other hand more specific than *head* and are therefore linked to it with an eq_hyperonym relation. In all three cases a direct-equivalent relation (eq_synonym) is expressed to new (non-English) ILI-records which are added (see Alonge 1996 and Climent et al 1996 for a list of all equivalence relations). This illustrates that the ILI is the superset of all concepts occurring in the other wordnets. The latter is mainly necessary to express equivalence relations between two non-English meanings, such as *dedo* and *dito*, even when there is no English equivalent.

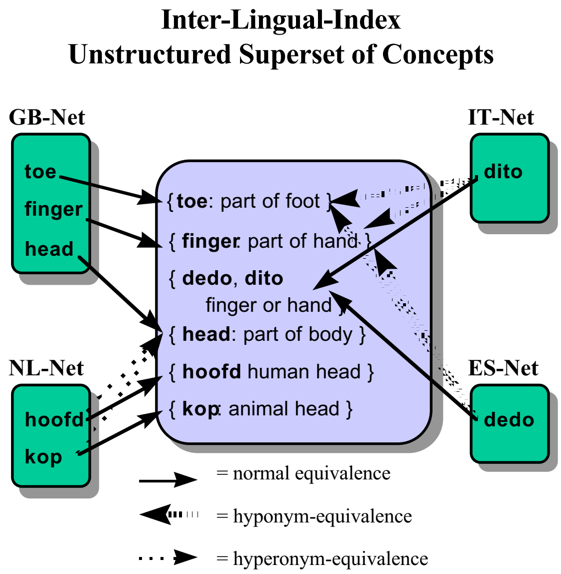

Figure 3.

Summarising, the design of the multilingual database has the following advantages:

〈 it will be possible to use the database for multilingual information retrieval, by expanding words in one language to related words in another language via the ILI.
〈 the different wordnets can be compared and checked cross-linguistically which will make them more compatible.
〈 language-dependent differences can be maintained in the individual wordnets.
〈 it will be possible to develop the wordnets at different sites relatively independently.
〈 language-independent information such as the domain-knowledge, the analytic top-concepts and information on instances can be stored only once and can be made available to all the language-specific modules via the inter-lingual relations.
〈 the database can be tailored to a user's needs by modifying the top-concepts, the domain labels or instances, (e.g. by adding semantic features) without having to know the separate languages or to access the language-specific wordnets.

Further details on the motivation and design of the database can be found in Díez et al (1996) and Bloksma et al (1996).

3. 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

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 see Alonge (1996) and Climent et al (1996).

3.1 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}
HAS_MERO_PART: c1 {door}
HAS_MERO_PART: c2d1 {jet engine}
HAS_MERO_PART: c2d2 {propeller}

{door}
HAS_HOLO_PART: d1 {car}
HAS_HOLO_PART: d2 {room}
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 seems to be of another order.

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.

3.2 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 synonymy and 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.

〈 in other languages it is either very difficult to distinguish nouns and verbs or the distinction does not even exist (Lyons 1977).

〈 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.

3.2 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):

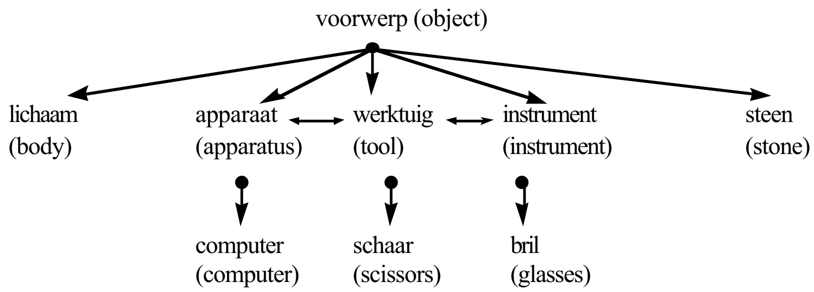

Figure 3.

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 speaker 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 replace 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.

3.3 SUBEVENT-relation

According to Fellbaum (Miller et al, 1990: 45) the entailment relation underlies all verbal relations: «the different relations that organize 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 the 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).

3.4 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.

3.6 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.

4. Conclusion

We have described a multilingual database which is designed in such a way that the language-specific wordnets can be stored as language-internal systems, while language-independent information can be shared via the so-called Inter-Lingual-Index or ILI. The ILI also embodies the links across the meanings of different languages. Users can then customise the database by modifying the Top-Concepts and Domain labels assigned to the ILI-records, without having knowledge of the different languages.

Furthermore, we have described the language-internal relations in so far as they are different from WordNet1.5 and relevant to the use for information-retrieval purposes. We have seen that especially the rigid separation of the meanings belonging to different parts-of-speech in WordNet1.5 is abandoned in EuroWordNet. This makes it possible to closely match meanings or expand to words that may have very different surface realisations within and across languages. The following general classes of relations are then distinguished:

Only within one Part-of-Speech:

Relation	*Example*
NEAR_SYNONYMY	apparatus N - machine N
HYPERONYMY/HYPONYMY	car N - vehicle N, move V - walk V
ANTONYMY	open V - close V
HOLONYMY/MERONYMY	head N - nose N

(Possibly) Cross-Part-of-Speech relations:

Relation	*Example*
XPOS_NEAR_SYNONYMY	dead A- death N; adorn V - adornment N
XPOS_HYPERONYMY/HYPONYMY	love V - emotion N
XPOS_ANTONYMY	live V - dead A
CAUSE	kill V - die V, die V - death N
SUBEVENT	buy V - payment N; sleep V - snore V
ROLE/INVOLVED	write V - pencil N; teach V - teacher N
STATE	poor N - poor N

In addition we have shown the use of labels and additional relations to differentiate the direction and strength of implications. This can also be used to develop different strategies for expanding words to related variants.

The database will contain the same generic vocabulary for the different languages, including the most frequent words in general corpora. It will be demonstrated in a multi-lingual retrieval task, where we will investigate the possibility to provide novice users access to data, e.g. via tree-browsing. In this respect, we see the usefulness of such a generic database in the added value to keyword-based retrieval. A first version of the database will be released at the end of 1997 and the final version will be tested in the IR-task at the end of 1998. The final results will be publicly available via licensing.

References

Alonge, Atonietta

1996 Definition of the links and subsets for verbs, EuroWordNet Project LE4003, Deliverable D006. University of Amsterdam, Amsterdam. Http: //www.let.uva.nl/~ewn.

Bloksma, L., P. Díez-Orzas, and P. Vossen,

1996 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.

Climent, Salvador, Horacio Rodríguez, Julio Gonzalo

1996 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.

Díez Orzas, P. , Louw M. and Forrest, Ph

1996 High level design of the EuroWordNet Database. EuroWordNet Project LE2-4003, Deliverable D007.

Miller G.A, R. Beckwidth, C. Fellbaum, D. Gross, and K.J. Miller

1990 «Introduction to WordNet: An On-line Lexical Database, in: International Journal of Lexicography, Vol 3, No.4 (winter 1990), 235-244.

Vossen, P.

1996 «Right or wrong: combining lexical resources in the EuroWordNet project», in M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L. Rogstrom, C.R. Papmehl, Proceedings of Euralex-96, Goetheborg, 1996, 715-728

Vossen, P. and A. Bon

1996 «Building a semantic hierarchy for the Sift project», Sift deliverable D20a, LRE 62030. Computer Centrum Letteren, University of Amsterdam.