Multilingual design of EuroWordNet

Piek Vossen, University of Amsterdam

Pedro Díez-Orzas, University of Madrid Alfonso X El Sabio

Wim Peters, University of Sheffield

Abstract

This paper discusses the design of the EuroWordNet database, in which semantic databases like WordNet1.5 for several languages are combined via a so-called inter-lingual-index. In this database, language-independent data is shared and language-specific properties are maintained as well. A special interface has been developed to compare the semantic configurations across languages and to track down differences. The pragmatic design of the database makes it possible to gather empirical evidence for a common cross-linguistic ontology.

1 Introduction

EuroWordNet is an EC-funded project (LE2-4003) that aims at building 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 (Miller90), i.e. synonyms are grouped in synsets, which in their turn are related by means of basic semantic relations.

The EuroWordNet database 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.

The multilingual nature of this conceptual database raises methodological issues for its design and development. First there is the question of which architecture to adopt. We have considered four possible designs:

a) Linking by pairs of languages.
b) Linking through an structured artificial language
c) Linking through one of the languages
d) Linking through an non-structured index

The first option (a) is to pair-wise link the languages involved. This makes it possible to precisely establish the specific equivalence relation across pairs of languages, but it also multiplies the work by the number of languages to be linked. Furthermore, the addition of a new language will ask for the addition of new equivalence relations to all the other languages, with all the possible consequences. The second option (b) is to link the languages through an structured language-neutral inter-lingua. A language-independent conceptual system or structure may be represented in an efficient and accurate way but the challenge and difficulty is to achieve such a meta-lexicon, capable of supplying a satisfactory conceptual backbone to all the languages. A drawback from a methodological point of view is that new words that are added in one of the languages might call for a revision of a part of the language-independent network.

As a third possibility the linking can be established through one of the languages. This resolves the inconveniences and difficulties of the former two options, but forces an excessive dependency on the lexical and conceptual structure of one of the languages involved. The last possibility (d) is to link through a non-structured list of concepts, which forms the superset of all concepts encountered in the different languages involved. This list does not satisfy any cognitive theory, because it is an unstructured index with unique identifiers for concepts that do not have any internal or language-independent structure. This has the advantage that it is not necessary to maintain a complex semantic structure that incorporates the complexity of all languages involved. Furthermore, the addition of a new language will minimally affect any of the existing wordnets or their equivalence relations to this index.

For pragmatic reasons we have chosen design (d). An unstructured index as a linking device is most beneficial with respect to the effort needed for the development, maintenance, future expansion and reusability of the multilingual database. Of course the adopted architecture is not without its difficulties. These are especially crucial in the process of handling the index and creating tools for the developers to obtain a satisfactory result.

Tasks such as identifying the right inter-lingual correspondence when a new synset is added in one language, or how to control the balance between the languages are good examples of issues that need to be resolved when this approach is taken.

In this paper we will further explain the design of the database incorporating the unstructured multilingual index. The structure of this paper is then as follows: first we will describe the general architecture of the database with the different modules. In section 3 we will discuss how language-specific relations and complex-equivalence relations are stored. Finally, section 4 deals with the specific options to compare the wordnets and derive information on the equivalence relations and the differences in wordnet structure.

2. High-level Design of the EuroWordNet Database

All language specific wordnets will be stored in a central lexical database system. Each wordnet represents a language-internal system of synsets with semantic relations such as hyponymy, meronymy, cause, roles (e.g. agent, patient, instrument, location). 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. The ILI starts off as an unstructured list of WordNet1.5 synsets, and will grow when new concepts will be added which are not present in WordNet1.5 (note that the actual internal organization of the synsets by means of semantic relations can still be recovered from the WordNet database which is linked to the index as any of the other wordnets). The only organization that will be provided to the ILI is via two separate ontologies which are linked to ILI records:

- the top-concept ontology: which is a hierarchy of language-independent concepts, reflecting explicit opposition relations (e.g. Object and Substance).
- a hierarchy of domains labels which relate concepts on the basis of scripts or topics, e.g. "sports", "water sports", "winter sports", "military", "hospital".

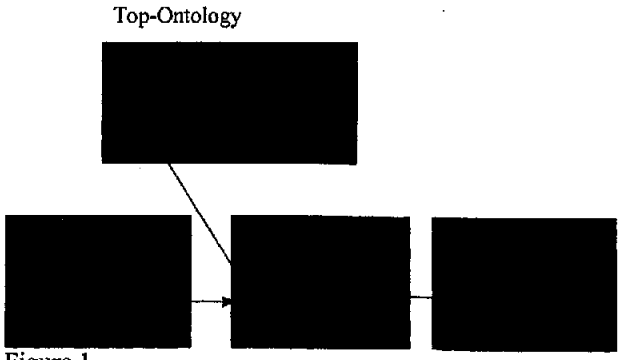

Figure 1.

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 top-concepts *Object* and *Substance*. The ILI-record *object* is linked to the Top-Concept *Object*. Since the Dutch synset *voorwerp* has an equivalence-relation to the ILI-record the Top-Concept *Object* also applies to the Dutch synset. Furthermore, it can be applied to all Dutch synsets related via the language-internal relations to the Dutch *voorwerp*.

Both hierarchies will enable a user to customize 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. Finally, we save space by storing the language-independent information only once.

The overall modular structure of the EuroWordNet database can then 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, the Domain Ontology and the Top-Concept Ontology.

Language internal Relationships		Language Module A	Language Module A
Interlingual relationships		Language Module A	ILI Module
Language Independent Relationships	Domain Internal Module Relationships	Domain Module	Domain Module
	Top-Concept Internal Module Relationships	Top-Concept Module	Top-Concept Module
	External Module Relationships	Domain Module Top-Concept Module	ILI Module

Table 1: Main categories of relationships

Three different types of relationships are necessary in this architecture, summarized in the table 1. The relationships operate upon five different types of data entities: Word-Meanings, Instances, ILI records, Domains and Top-Concepts. The Word-Meanings are senses with denotational meanings (*man*) while the Instances are senses with referential meanings (*John Smith*).

Figure 2 gives a simplified overview of how the different modules are interconnected. In the middle the ILI is given in the form of a list of ILI-records: "animal", "mammal", ... "mane", "Bob", with relations to the language-modules, the domains, and the top-concepts. Two examples of inter-linked domains (D) and top-concepts (TC) are given above the ILI-records. The boxes with language-names (Spanish, English, Dutch, Italian and WN1.5) represent the Language Modules and are centered around the ILI. For space limitations, we only show a more detailed box for the Spanish module. In this box we see examples of hyponymy and meronymy relations between Spanish word-meanings and some of the equivalence-relations with the ILI-records. The full list of relations distinguished, its characteristics and assignment tests, as well as the structures of the different records can be found in the EuroWordNet deliverables D005, D006, D007 (available at: http://www.let.uva.nl/~ewn).

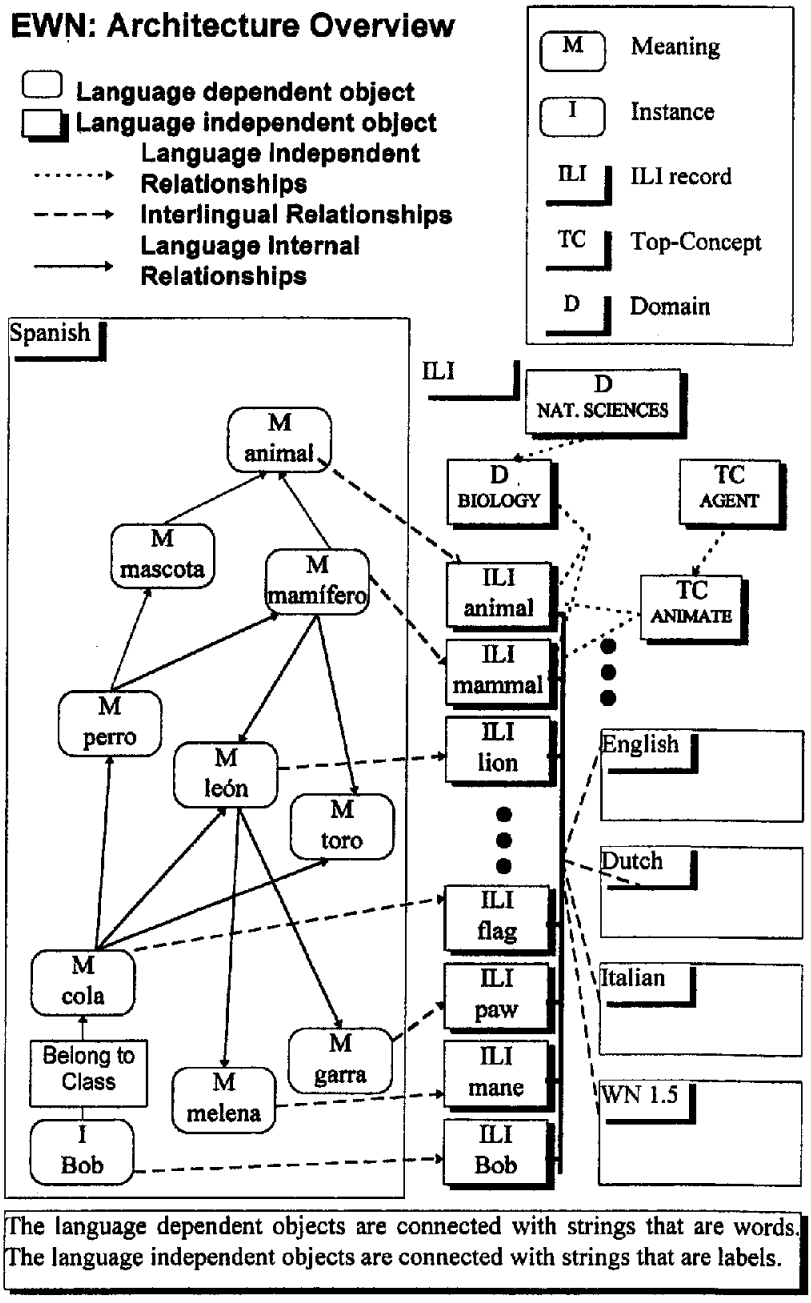

Figure 2

Next to the language-internal relations there are also 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: e.g. Dutch *hoofd* only refers to **human head** and *kop* only refers to **animal head**, while English has *head* for both.
- HAS_EQ_HYPONYM when a meaning can only be linked to more specific ILI-records: e.g. Spanish *dedo* which can be used to refer to both *finger* and *toe*.

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.

As mentioned above, the ILI should be the super-set 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-Spanish) to the same ILI-record even when there is no English or WordNet equivalent.
- it should be possible to store 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 all WordNet1.5 synsets but eventually it will be updated with language-specific concepts using a specific update policy:

- a site that cannot find a proper equivalent among the available ILI-concepts will link the meaning to another ILI-record using a so-called complex-equivalence relation and will generate a potential new ILI-record (see table 2).
- after a building-phase all potentially-new ILI-records are collected and verified for overlap by one site.
- a proposal for updating the ILI is distributed to all sites and has to be verified.
- the ILI is updated and all sites have to reconsider the equivalence relations for all meanings that can potentially be linked to the new ILI-records.

3. Mismatches and language-specific semantic configurations

Within the EuroWordNet database, the wordnets can be compared with respect to the language-internal relations (their lexical semantic configuration) and in terms of their equivalence relations. The following general situations can then occur (Vossen 1996).

1. a set of word-meanings across languages have a simple-equivalence relation and they have parallel language-internal semantic relations.
2. a set of word-meanings across languages have a simple-equivalence relation but they have diverging language-internal semantic relations.
3. a set of word-meanings across languages have complex-equivalence relations but they have parallel language-internal semantic relations.
4. a set of word-meanings across languages have complex-equivalence relation and they have diverging language-internal semantic relations.

Dutch	hoofd	human head	HAS_EQ_HYPERONYM	head
Dutch	kop	animal head	HAS_EQ_HYPERONYM	head
Spanish	dedo	finger or toe	HAS_EQ_HYPONYM	finger
Spanish	dedo	finger or toe	HAS_EQ_HYPONYM	toe

Table 2: Complex-equivalence relations for mismatching meanings.

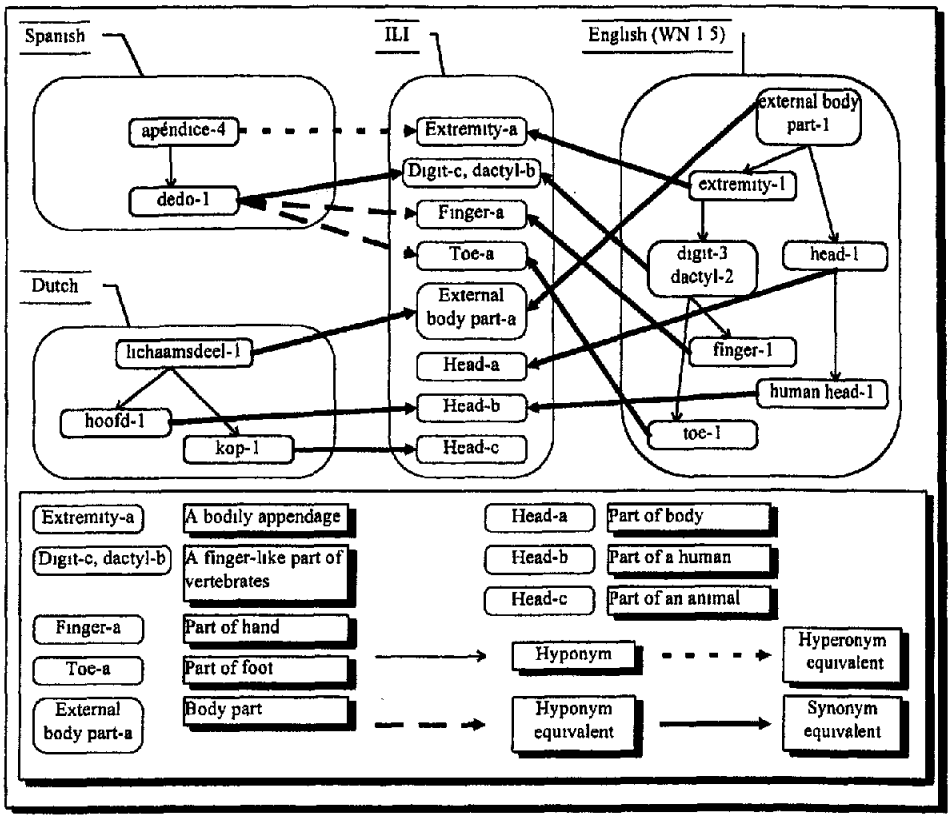

Figure 3.

Figure 3[1] gives some examples of the different mismatches. Here we see that *head-1* represents an intermediate level between *human-head-1* and *external-body part-1* in WordNet1.5 which is missing between their Dutch equivalent *lichaamsdeel-1* and *hoofd-1*. While the equivalence relations match, the hyponymy-structure does not (situation 2 above). Furthermore, *kop-1* does not match any synset in WordNet1.5. In the Spanish-English example we see on the other hand that *apéndice-4* and *dedo-1* have complex equivalence relations which are not incompatible with the structure of the language-internal relations in the Spanish wordnet and in WordNet1.5 (situation 4 above).

In general we can state that situation (1) is the ideal case. In the case of (4), it may still be that the wordnets exhibit language-specific differences which have lead to similar differences in the equivalence relations. Situation (2) may indicate a mistake or it may indicate that equivalent meanings have been encoded in an alternative way in terms of the language-internal relations. Situation (3) may also indicate a mistake or it may be the case that the meanings are non-equivalent and therefore show different language-internal configurations.

4. Comparing the wordnets via the multilingual index

The EuroWordNet database is developed in tandem with the Novell ConceptNet toolkit (Díez-Orzas et al 1995). This toolkit makes it possible to directly edit and add relations in the wordnets. It is also possible to formulate complex queries in which any piece of information is combined. Furthermore, the ConceptNet toolkit makes it possible to visualize the semantic relations as a tree-structure which can directly be edited. These trees can be expanded and shrunk by clicking on word-meanings and by specifying so-called filters indicating the kind and depth of relations that need to be shown.

However, to get to grips with the multi-linguality of the database we have developed a specific interface to deal with the different matching problems. The multilingual interface has the following objectives:

- it should offer new or better equivalence relations for a set of word-meanings
- it should offer better or alternative language-internal configurations for a set of word-meanings
- it should highlight ill-formed configurations
- it should highlight ill-formed equivalence relations

[1] Obviously, the correspondence between WordNet and the ILI is very high, because it is built from it. Only in later stages of development, new ILI records occur

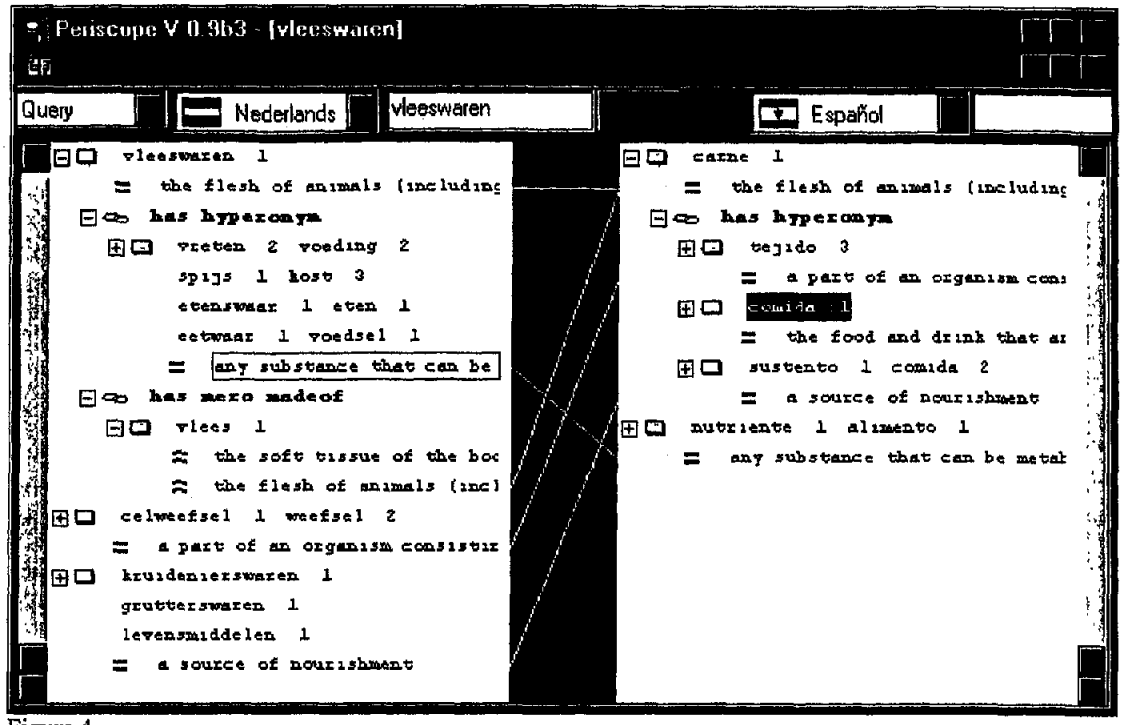

Figure 4.

For visualising these aspects we designed an interface in which two wordnets can be aligned (see Cuypers and Adriaens 1997 for further details). In the screen-dump of the interface (figure 4) we see a fragment of the Dutch wordnet in the left box and a fragment of the Spanish wordnet in the right box.[2] The dark squares represent the meanings (WMs) in the languages which are interconnected by lines labeled with the relation type that holds: has_hyperonym, has_mero_madeof. Each meaning is followed by the synset (as a list of variants with a sense-number) and on the next lines by the ILI-records to which it is linked (if any). These ILI-records are represented by their gloss (here all taken from WordNet1.5) and the kind of equivalence relation is indicated by a preceding icon, = for EQ_SYNONYM and ≈ for EQ_NEAR_SYNONYM. By displaying the wordnets adjacently and by specifying the ILI-records separately for each synset in each tree the matching of the ILI-records can be indicated by drawing lines between the same ILI-records. When comparing wordnets one specific language can be taken as a starting point. This language will be the Source Language (SL). The SL is compared with one or more other languages which will be called the Reference Languages (RLs).

There are then two general ways in which the aligned wordnets can be accessed:

- given a (set of) WM(s) in a source wordnet with their corresponding ILIR(s), generate the same ILIRs in the adjacent wordnet box with the corresponding WMs in the reference wordnet.
- given two comparable wordnet structures visualise the matching of the ILIRs: i.e. draw the lines between the ILI-records that are the same.

In the first option, a WM is first 'translated' into the second wordnet box, yielding a parallel twin-structure of ILI-records. Next the language-specific configuration of the Reference-wordnet can be generated (bottom-up). This gives you the semantic structuring of a particular set of WMs according to another wordnet as compared to the Source-wordnet.

In the second option the structures of both the Reference and the Source wordnet are compatible and the inter-lingual relations are compared relative to this structure. Each set of ILI-records represents the most direct matching of a fragment of a wordnet from the available fund of ILI-records, regardless of the matching of the other wordnet. The equivalence relations of these compatible fragments can then directly be compared. Loose-ends at either site of the ILI-records can be used to detect possible ILIR-records that have not been considered as translations in one wordnet but have been used in another wordnet. Differences in the kind of equivalence relations of WMs with compatible structure are suspect. Obviously, a comparison in this way only makes sense if the semantic-scope of the language internal relations is more or less the same.

Both these options are illustrated in the above screen-dump. For example, the Dutch *vleeswaren:1* (meat-products) has an EQ_SYNONYM relation with *meat.2* (= the flesh of animals ...), where the sense

[2] Only part of the available information is shown in this screen-dump.

numbers do not necessarily correspond with WordNet1.5 numbers, and a HAS_HYPERONYM relation to the synset *voedsel:1*. The latter is in its turn linked to the ILI-synset *food:1*(=any substance that can be metabolized...). We then copied the ILI-record *meat 2* into the Spanish wordnet yielding *carne 1* as the synset linked to it. By expanding the hyperonymy-relations for *carne·1* we see that the Spanish wordnet gives three hyperonyms: *tejido·3* (tissue:1 = a part of an organism ..), *comida:1* (fare:1 = the food and drink that are regularly consumed), and *sustento 1* (nourishment:1 = a source of nourishment), all linked to ILI-records different from the Dutch case. When generating back the matching Dutch synsets for these hyperonyms it becomes clear that they are all present in this fragment, except for *comida·1* (fare:1) which does not yield a corresponding Dutch synset. First of all this comparison gives us new hyperonyms that can be considered and, secondly, it gives us a new potential ILI-record *fare:1* for the Dutch wordnet. Further expanding the Dutch wordnet also shows that there is a closely-related concept *vlees:1* (the stuff where meat-products consist of) which matches both *meat.2* and *flesh:1*(= the soft tissue of the body...). This concept thus partially matches the Spanish *carne:1*. Since there is no matching Spanish concept related to *flesh 1* the Dutch wordnet thus in its turn suggests a new potential ILI-record for the Spanish wordnet. In this way the aligned wordnets can be used to help each other and derive a more compatible and consistent structure.

Given the fact that we allow for a large number of language-internal relations and six types of equivalence relations, it may be clear that the different combinations of mismatches is exponential. Therefore we are differentiating the degree of compatibility of the different mismatches: some mismatches are more serious than others. First of all, some relations in EuroWordNet have deliberately been defined to give somewhat more flexibility in assigning relations. In addition to the strict synonymy-relation which holds between synset-variants there is also the possibility 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*. Despite the tests for each relation there are always border-cases where intuitions will vary. Therefore it makes sense to allow for mismatches across wordnets where the same type of equivalence relation holds between a single synset in one language and several synsets with a NEAR_SYNONYM relation in another language.

As we have seen above, a single WM may be linked to multiple ILI-records and a single ILI-record may be linked to multiple WMs. This allows for some constrained flexibility. The former case is only allowed when another more-global relation EQ_NEAR_SYNONYM has been used (see above). In the reverse case, the same ILI-record is either linked to synsets which have a NEAR_SYNONYM relation among them (in which case they can be linked as EQ_SYNONYM or as EQ_NEAR_SYNONYM of the same ILI-record) or any other complex equivalence relation which parallels the relation between the WMs. Thus, two WMs which have a hyponymy-relation among them and which are linked to the same ILI-record should have equivalence-relations that parallel the hyponymy-relation: EQ_HAS_HYPERONYM and EQ_SYNONYM. A final type of flexibility is built in by distinguishing subtypes of relations. In addition to more specific meronymy-relations such as member-group, portion-substance there is an a-specific meronymy relation which is compatible with all the specific subtypes.

In addition to more global or flexible relations, we also try explicitly define compatibility of configurations. First of all, differences in levels of generality are acceptable, although deeper hierarchies are preferred. So if one wordnet links *dog* to *animal* and another wordnet links it to *mammal* and only via the latter to *animal* first these structures are not considered as serious mismatches. Furthermore, since we allow for multiple hyperonyms it is possible that different hyperonyms may still both be valid. To make the compatibility of hyperonyms more explicit, the most frequent hyperonyms can be defined as allowable or non-allowable combinations. For example, a frequent combination such as *act* or *result* can be seen as incompatible (and therefore have to be split into different synsets), whereas *object* or *artifact* are very common combinations.

Finally, we have experienced that some relations tend to overlap for unclear cases. For example, intuitions appear to vary on causation or hyponymy as the relation between Dutch pairs such as *dichttrekken* (close by pulling) and *dichtgaan* (become closed). In these cases it is not clear whether we are dealing with different events in which one causes the other or one makes up the other. The events are fully co-extensive in time: there is no time point where one event takes place and the other event does not. This makes them less typical examples of cause-relations. By documenting such border-line cases we hope to achieve consensus about the ways in which they should be treated and the severity of the incompatibility.

5. Conclusion

The multilingual EuroWordNet database thus consists of separate language-internal modules, separate language-external modules and an inter-lingual module which has the following advantages:

- it will be possible to use the database for multilingual retrieval.
- 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.
- 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.

At the same time, the fact that the Inter-Lingual-Index or ILI is unstructured has the following major advantages:

- 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 structure of the database and the strategies for its implementation have been chosen out of pragmatic considerations. The architecture will allow maximum efficiency for simultaneous multilingual implementation in more than one site, and will offer an empirical view on the problems related to the creation of an inter-lingua by aligning the wordnets, thus revealing mismatches between 'equivalent' semantic configurations. These mismatches may be due to:

- a mistake in the equivalence-relations (inter-lingual links)
- a mistake in the Language Internal Relations
- a language-specific difference in lexicalization

By using the cross language comparison and the tools described in section 4 a particular series of mismatches can provide criteria for selecting that part of the semantic network which needs inspection, and may give clues on how to unify diverging semantic configurations. This will constitute the first step towards generating an interlingua on the basis of a set of aligned language-specific semantic networks.

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.

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

Díez-Orzas P. and I. Cuypers,
1995 *The Novell ConceptNet*, Internal Report, Novell Belgium NV.

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, [also available as EuroWordNet-working paper at Http: //www.let.uva.nl/~ewn].

Vossen, P.
1997 EuroWordNet: a multilingual database for information retrieval, in: Proceedings of the Delos workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich, [also available as EuroWordNet-working paper at Http: //www.let.uva.nl/~ewn].