Title: The Multilingual design of the EuroWordNet Database

Authors:

Piek Vossen,
University of Amsterdam,
Spuistraat 134, 1012 VB Amsterdam, The Netherlands.

Wim Peters,
University of Sheffield,
211 Portobello Street, S1 4 DP Sheffield, Great Brittain

Pedro Díez-Orzas,
University of Madrid Alfonso X El Sabio,
Avenida de la Universidad, 1
28691 Villanueva de la Caqada Madrid, Spain

Abstract

This paper discusses the design of the EuroWordNet database, in which semantic databases like WordNet1.5 for several languages are combined via an inter-lingua. 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.

Published in:
Kavi Mahesh (ed.) Ontologies and multilingual NLP, Proceedings of workshop at IJCAI-97, Nagoya, Japan, August 23-29.

The Multilingual design of the EuroWordNet Database

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[1] 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 [Miller et al. 1991], 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 a structured artificial language
c) Linking through one of the languages
d) Linking through a 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 a 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

[1] EuroWordNet is a joint project of Amsterdam University (co-ordinator), the Istituto di Linguistica Computazionale del CNR (Pisa), the Fundacíon Universidad Empresa (Barcelona University, UNED-Madrid and the Politechnica de Catalunya), Sheffield University and Novell Linguistic Development (Antwerp).

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 reports on the building of wordnet fragments and basic ontologies for the first subset of Basic Concepts. These Base Concepts are selected for their key-role in the wordnet-structures.

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 (see Figure 1). 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 domain labels, 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 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*. A similar inference can be made for all Italian meanings linked to *ogetto*.

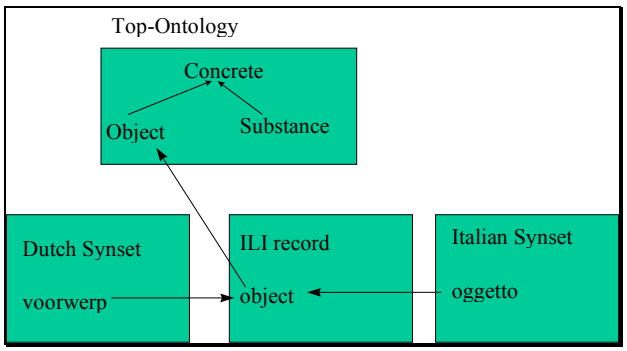

Figure 1.

The top-concept ontology and the domain-ontology 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. Figure 2 gives a simplified overview of how the different modules are interconnected (see [Díez_Orzas et al. 1996] and [Bloksma et al. 1996] for further details). 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. Further

information can be found at: http: //www.let.uva.nl/~ewn).

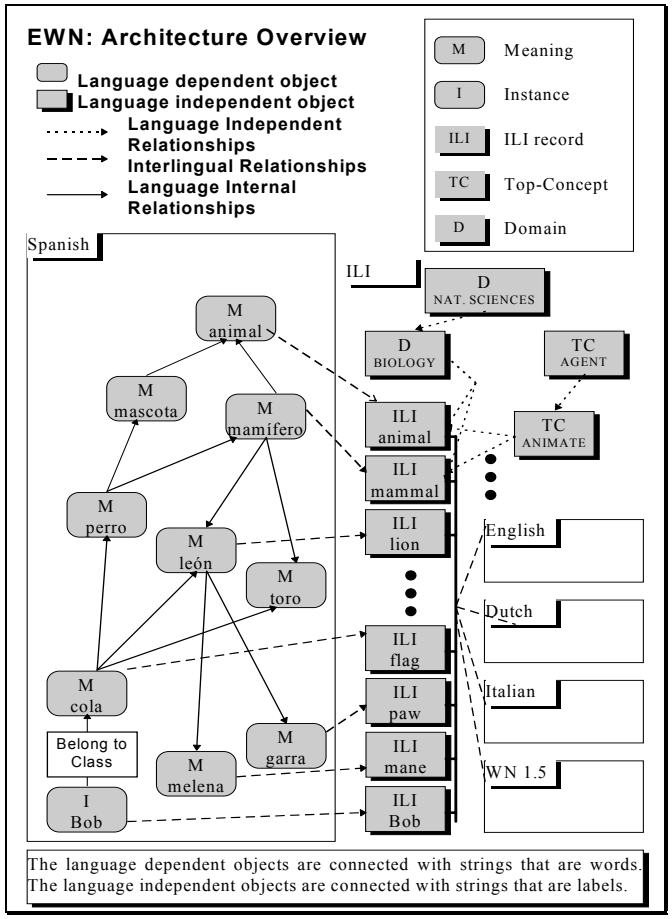

Figure 2.

Next to the language-internal relations there are also six different types of inter-lingual relations. The most straightforward relation is EQ_SYNONYM, which applies to meanings 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 *dito* to Spanish *dedo*) 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 1).
- 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.

New ILI-synset	*New ILI-gloss*	*Equivalence relation*	*Target-concept*
hoofd	human head	has_eq_hyperonym	head
kop	animal head	has_eq_hyperonym	head
dedo	finger or toe	has_eq_hyponym	finger
dedo	finger or toe	has_eq_hyponym	toe

Table 1.

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

Figure 3 gives some examples of the different mismatches. Here we see how *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.

Given the large number of language-internal relations and six types of equivalence relations the different combinations of mismatches is exponential. We therefore differentiate 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*. Consequently, 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 are allowed.

As we have seen above, a single word-meaning (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 from a single WM to multiple ILI-records. For example, in the Dutch resource there is only one sense for *schoonmaken* (to clean) which simulatenously matches with at least 4 senses of *clean* in WordNet1.5:

- {making clean by removing filth, or unwanted substances}
- {remove unwanted substances from, such as feathers or pits, as of chickens or fruit}
- {remove in making clean}
- {remove unwanted substances from – (as in chemistry)}

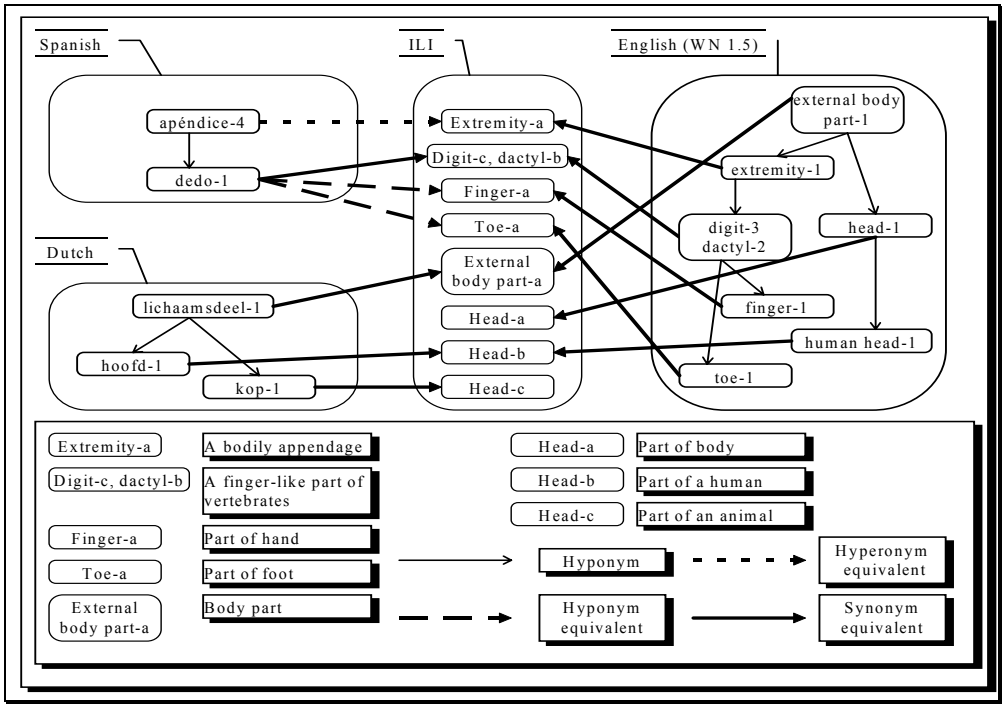

Figure 3.

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 to the same ILI-record) or any other complex equivalence relation which parallels the relation between the WMs.

Thus, two WMs (e.g. Dutch *jenever* (gin) and *citroenjenever* (a specific type of gin)) which have a hyponymy-relation among them and which are linked to the same ILI-record (e.g. *gin*) should have equivalence-relations that parallel the hyponymy-relation: HAS_EQ_HYPERONYM and EQ_SYNONYM respectively. 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.

Next to more global or flexible relations, we also try to 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 *dog* to *mammal* and *mammal* to *animal,* this is not considered as a serious mismatch. 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. Incompatibility of major concepts can be expressed by linking them to disjoint top-concepts (see below).

Finally, specific mismatches to WordNet1.5 or across the wordnets can be investigated and manually revised using the multilingual EuroWordNet viewer: Periscope [Cuypers and Adriaens 1997]. This viewer has specificallly been designed to align wordnet tree-structures which are matched via the ILI. The major functionality of the Periscope-viewer is to:

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

4. Building the wordnets

The challenge for EuroWordNet is how we can both provide the flexibility to develop the resources independently at different sites (using different resources and tools as a starting point and maintaining language-specific properties), and, on the same time, make sure that the coverage and content are still compatible.

One of the measures to ensure compatibility[2] of the wordnets is the definition of a set of **Common Base Concepts** that cover the major concepts of the wordnets. The Base-Concepts have first been defined for each language-resource using two major criteria:

- having many relations with other meanings (e.g. the highest frequency as a genus word in dictionary definitions).
- having a high position in the hierarchies (e.g. any meaning which is used to classify one of the previous meanings).

Next each site has translated the local Base-Concepts to the closest WordNet1.5 synsets. Those translations (WordNet1.5 synsets) chosen by at least two other sites have been included in the set of Common Base Concepts.[3] This has resulted in a set of 871 WN1.5-synsets (694 nouns and 177 verbs). Only 30 synsets of these have been selected by all sites (24 noun synsets, 6 verb synsets). Each site has then extended their local selection of BCs with equivalences (or most close concepts) for the Common BCs not yet covered. The production of the local wordnets proceeds top-down by first encoding the direct semantic context for the BCs and next extending the hierarchies in depth.

These Base-Concepts (BCs) not only form the core of each wordnet but also tend to be very polysemous, to have poor and vague definitions, many synonyms, and complex morpho-syntactic properties. By making a distinction between BCs and more-specific meanings we will be able to focus the manual work on the former and to apply the (semi-)automatic techniques to the latter.

To get to grips with these Base Concepts they have been globally clustered into 79 semantic classes. These classes are organized in the form of a preliminary top-ontology: see Figure 4. The top-ontology is based on top-nodes in WordNet1.5, ontologies from other EC-projects (Acquilex and Sift) and Aktions-Art models. Furthermore, the ontology was adapted to represent the variety of concepts in the set of Common Base Concepts. The first division in the ontology is made between First-Order-Entities and High-Order-Entities, where the former are concrete objects and substances and the latter events, processes, relations, properties and states. Note that in EuroWordNet both nouns and verbs can be linked to the same classifications and top-concepts. Furthermore, each BC may belong to any number of Top-Concepts, e.g. some of the concrete nouns are classified in terms of their Origin, Form, Composition and Function [compare Pustejovsky 1991].

[2] Other measures are: loading the data in a common database; the use of common tests in each language to verify the semantic relations; the development of user-guides for building the wordnets.

[3] Special measures have been taken to prevent that different but closely-related senses have been chosen by different sites. This is very likely to happen because the Base Concepts tend to very polysemous and are poorly defined. Furthermore, we have inspected all the rejected concepts to see whether they represent new concepts not represented by the common selection.

Figure 4: Ontology of Top-Concepts

The clustering of common BCs per Top-Concept has been verified by each site for the local wordnets by applying it to their equivalences. Both the clustering and the top-concept ontology has thus been revised in several circles. The production of wordnet fragments is further monitored by, exchanging major problems and solutions per Top-Concept cluster.

As discussed above, it is possible to find potential areas for revision by comparing a wordnet with the other wordnets (including WordNet1.5). These revisions can be done by each site individually but also by revising the matching of the ILI-records with the top-concepts and by revising the ILI-records as such. One of the major measures we envisage is to globalize the matching of meanings to the ILI-synsets (mainly WordNet1.5 sense-distinctions). Typically, many mismatches have to do with differences in the sense-differentiation across the resources. Especially in WordNet1.5 there appears to be over-differentiation of senses for specific meanings which are often represented by a single meaning in traditional resources (see above: Dutch *schoonmaken* and WordNet *clean*). Instead of keeping the extremely-differentiated meanings one global ILI-record would suffice (which will still be linked to the more specific meanings in WordNet1.5). A more coarse differentiation of senses also minimalizes the danger that equivalences across the wordnets are related to different senses of the same word in the ILI.

Another typical mismatching problem has to do with the inconsistent representation of regular patterns of polysemy e.g.: *church* may be defined as a *building*, the *service* or both. This problem could be solved in a similar way by producing relations between classes of potential sense-extension, such as animal-meat, building-institute [Hamp and Feldweg 1997]. A language-specific WM linked to *church-building* can thus automatically be linked to another WM linked to *church-service,* even though none of the wordnets has both senses.

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.
- language-dependent differences can be maintained in the individual wordnets.
- language-independent information (the domain-labels, top-concepts, instances) is 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, 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. There is 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 be seen as a fund of concepts 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. 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

[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.

[Cuyper and Adriaens 1996] Cuypers, I. And G. Adriaens,. "Periscope: the EWN Viewer", EuroWordNet Project LE4003, Deliverable D008d012. University of Amsterdam.

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

[Hamp and Feldweg 1997] Hamp, B., H. Feldweg, "GermaNet: a Lexical-Semantic Net for German", in: Vossen, Calzolari, Adriaens, Sanfilippo, Wilks (eds.) Proceedings of the ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid.

[Miller *et al.* 1991] Miller G.A, R. Beckwidth, C. Fellbaum, D. Gross, and K.J. Miller , "Introduction to WordNet: An On-line Lexical Database, in: International Journal of Lexicography, Vol 3, No.4 (winter 1990), 235-244.

[Pustejovsky 1991] Pustejovsky, J. "The Generative Lexicon", in: Computational Linguistics, 17, 4, 1991, Cambrdige MA, MIT Press, 409-442.

[Vossen 1996] Vossen, P. "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.