Towards a Universal Index of Meaning

Piek Vossen
Univ. of Amsterdam
The Netherlands
Piek Vossen@hum uva nl

Wim Peters
Univ. of Sheffield
U.K.
W Peters@dcs shef ac uk

Julio Gonzalo
UNED
Spain
julio@ieec uned es

Abstract

The Inter-Lingual-Index (ILI) in the EuroWordNet architecture is an initially unstructured fund of concepts which functions as the link between the various language wordnets The ILI concepts originate from WordNet1 5, and have been restructured on the basis of aspects of the internal structure of WordNet, links between WordNet and other resources, and multilingual mapping between the wordnets This leads to a differentiation of the status of ILI concepts, a reduction of the Wordnet polysemy, and a greater connectivity between the wordnets The restructured ILI represents the first step towards a standardized set of word meanings, is a working platform for further development and testing, and can be put to use in NLP tasks such as (multilingual) information retrieval

1 Introduction

EuroWordNet (LE2-4003, LE4-8328) develops a multilingual database with wordnets for 8 different European languages English, Dutch, Spanish, Italian, German, French, Czech and Estonian Further collaborations have been established with wordnet builders for Portuguese, Swedish, Basque, Catalan, Russian, Greek and Danish, who work according to the EuroWordNet specifications Each of the wordnets is structured as the Princeton Wordnet (Fellbaum, 1998) in terms of sets of synonymous words or so-called synsets between which basic semantic relations are expressed The synsets are based on the lexicalizations and expressions in each language Each wordnet therefore can be seen as a unique language-specific structure

In addition to the relations between synsets there is also a relation to a so-called Inter-Lingual-Index This Inter-Lingual-Index (ILI) is an unstructured fund of concepts, so-called ILI-records, with the sole purpose of linking synsets across languages Synsets that are linked to the same ILI-record can be said to be equivalent across two languages By means of the ILI it is thus possible to go from one wordnet to the other and to compare the lexicalization patterns across languages

The characteristics of the ILI are defined by its function to provide an efficient mapping across the meanings in the wordnets for the different languages Two major requirements follow from this

- the ILI should have a certain level of granularity,
- the ILI should be the superset of concepts that occur across languages

The first requirement is necessary to make the linking of meanings easier If many specialized meanings and interpretations are given it is more difficult to find mappings from a language-specific wordnet to the index The second requirement is necessary to be able to express an equivalence relation across synsets in two wordnets for which there is no equivalent in other wordnets

Initially, the ILI has been based on WordNet1 5 It is however a well-known problem that sense-differentiation is very inconsistent within and across resources including WordNet1 5 On the basis of the above criteria and by comparing the sense-differentiation across the wordnets we have therefore begun to adapt the ILI Four major revisions of the ILI are derived from these

- grouping sense-differentiations between which there is a systematic polysemy relation e g metonymy,
- grouping sense-differentiations that can be represented by more general sense-group
- adding sense-differentiations or concepts that occur in two wordnets but not in WordNet1 5
- differentiating the status of the ILI-records in terms of universality, productivity, and exhaustive linking

The sense-groupings lead to a coarser differentiation of senses which will make the ILI more effective for mapping senses across languages Furthermore, the differentiation of the status of ILI-records can be used to determine the relevance of

	Nouns			Verbs		
	Total	62780	32520	Total	12215	7455
		⋃{WN/IT/NL/ES}	⋃{IT/NL/ES}		⋃{WN/IT/NL/ES}	⋃{IT/NL/ES}
ES	24153	38,5%	74,3%	4074	33,4%	54,6%
IT	13950	22,2%	42,9%	3569	29,2%	47,9%
NL	20877	33,3%	64,2%	5562	45,5%	74,6%
⋂{ES/IT}	10449	16,6%	32,1%	2030	16,6%	27,2%
⋂{ES/NL}	14302	22,8%	44,0%	2778	22,7%	37,3%
⋂{IT/NL}	9445	15,0%	29,0%	2574	21,1%	34,5%
⋂{ES/IT/NL}	7736	12,3%	23,8%	1632	13.4%	21.9%

Table 1 Intersections of ILI references in English (WN) Dutch (NL), Spanish (ES) and Italian (IT)

finding a mapping to particular senses Eventually, the restructuring will result in a more universal list of sense-distinctions that can also be used for sharing NLP technology across languages, as a gold-standard in Word-Sense-Disambiguation (WSD) and for the testing WSD techniques across languages in (ROMAN)SENSEVAL (where similar sense-mapping problems have been encountered)

In this paper we discuss the restructuring of WordNet1 5 and the differentiation of the ILI-records derived from it along the above lines In section 2, we give an overview of the mapping of meanings in the wordnets that are currently available Section 3 gives an overview of the criteria that have been used to group closely related ILI-records, both on internal structural properties of WordNet1 5 and on the basis of cross-linguistic evidence Figures on the resulting increase of matching across the wordnets are given Section 4 describes the opposite restructuring Synsets that could not be linked to the ILI have been inspected to see how much overlap there is and what the status is of these concepts Finally, section 5 describes how the ILI can be used as a standardized set of concepts for NLP tasks for different languages and across languages

2 The Universality of meanings across wordnets

The wordnets in EuroWordNet are based on existing dictionaries and sense-inventories, where selections have been tested for corpus frequency (at least all more frequent words) and generality (at least all generic word meanings) As a multilingual database with a sense-based mapping EuroWordNet thus provides a unique possibility to find out how universal word senses are across languages on a large scale Currently, final figures are available for the Dutch, Italian and Spanish wordnets The size of each wordnet is between 30 and 45K synsets For comparison, WordNet1 5 has a size of about 80K synsets for nouns and verbs The synsets in these languages have been translated to the closest WordNet1 5 synset or ILI-record, using bilingual dictionaries, automatic mapping heuristics (Agirre and Rigau, 1996) and manual selection procedures (about 50% is checked manually) Not all synsets have an equivalence relation to the ILI, e g in the case of the Dutch wordnet 16% of the nouns and 11% of the verbs have no equivalence link In other cases different synsets refer to the same ILI-record or single synsets are linked to multiple ILI-records The number of ILI-record references in a wordnet therefore only weakly correlates with the actual size of the wordnet In Table 1, an overview of the number of ILI-records referred to in each wordnet and the intersection between them is given The figures are differentiated for nouns and verbs, where separate rows are given for each wordnet separately and the intersection of 2 and 3 wordnets The first column then gives the absolute numbers, the second column gives the percentage of all ILI-records occurring in all 4 resources (including WordNet1 5) the third column gives the percentage of the ILI-references occurring in the Spanish Italian and Dutch wordnet only

Without restructuring the ILI (see next section) we see that the intersection for nouns between wordnet pairs ranges between 30 and 44% of the total union of ILI-records occurring in all 3 wordnets Including WordNet1 5, the intersection goes down to 15 to 23% This lower coverage is obvious because the total union of the 3 languages is about 50% of WordNet1 5 In the case of verbs, we get similar results 27 to 37% intersection between wordnet pairs compared to the union of 3 languages and 16 to 23% if we also include WordNet1 5 (maximum coverage is 50%) The intersection of 3 languages is lower but close to the lowest intersection between language pairs 24% for nouns and 22% for verbs (out of the union of 3 languages) This corresponds with a set of 7736 nominal and 1632 verbal concepts that are (somehow) lexicalized in 4 languages The union of concepts lexicalized in 3 languages is of 18724 nouns and 4118 verbs

The wordnets for French, German, Czech and Es-

	Nouns	3 lang		4 lang		Verbs	3 lang		4 lang	
		18724		7736			4118		1632	
DE	4480	3366	75,1%	2085	46,5%	1959	1401	71,5%	771	39,4%
FR	5523	4147	75,1%	2602	47,1%	2534	1507	59,5%	770	30,4%
EE	2596	2100	80,9%	1428	55,0%	489	413	84,5%	284	58,1%
CZ	6754	5121	75,8%	2872	42,5%	1306	861	65,9%	474	36,3%

Table 2 Overlap of ILI references in German (DE), French (FR), Czech (CZ) and Estonian (EE) with the union of concepts lexicalized in three and four languages out of English, Dutch, Spanish and Italian

tonian are still under development However, core wordnets for the most important meanings have been finished varying from 3 to 10K synsets in size We can use this set to evaluate the shared set of meanings extracted for Dutch, Spanish and Italian Table 2 first gives the number of ILI-references for nouns and verbs, and in the next columns the intersection of these references with the ILI-records lexicalized in ,3 of the above languages and in 4 of the above languages

For nouns we see that 75 up to 85% of the nominal synsets and 60 to 85% of the verbal synsets are covered by the set occurring in at least 3 languages This means that the set of concepts occurring in at least 4 languages can be extended considerably The intersection with at least 4 languages, ranges from 42 to 55% for nouns and 30 to 58% for verbs

The high overlap of the relatively small wordnets is partly due to the common approach for building the wordnets, where each site develops the resources top-down starting from common set of 1300 Base Concepts Nevertheless, we can also expect that these selections cover many of the more general and frequent words that are polysemous, which cause most problems for WSD and linking meanings across languages

As such the core intersection is still valuable It can be used to derive an initial standardized set of core meanings that not only functions as an index in EuroWordNet but can also be used for developing a gold-standard for sense-tagging, for WSD and information retrieval, both monolingual and crosslingual Eventually the core intersection can be further condensed to a set of semantic tags Absence of a semantic tag set currently makes WSD fundamentally different from morphological disambiguation or tagging techniques (Wilks, 1998) If simple tagging techniques can be applied to large corpora (uniformly across languages) this information could be used to derive statistical information on the usage of an initial set of word meanings (possibly in different languages) Information on usage could then be used to further standardize the set of word meanings

It will be clear that the above measurements depart from WordNet1 5 as a standardized set There are two biases that may follow from this First of all the cross-lingual mapping of synsets or word senses may be improved if inconsistent sense-differentiation is somehow dealt with Secondly, a universal list can not just be based on English We thus have to consider the status of synsets in the other languages that could not be matched with WordNet1 5 synsets Both aspects will discussed in the next two sections

3 Restructuring the ILI

Sense distinctions in Wordnet1 5 are often too fine-grained for WSD purposes which makes it difficult to link wordnets for polysemous words Also the systematic relatedness between word senses has not been made explicit in WordNet The clustering of WordNet derived concepts into larger conceptual chunks that represent meaning at a higher or more underspecified level of semantic description enhances the interconnectivity of wordnets and can be be put to use in NLP applications such as Information retrieval

We have distinguished two types of these clusters which differ in their semantic characteristics They are *metonymy* and *generalization* and will be discussed in the following subsections

3 1 Metonymy

Metonymy can be defined as a (semi-)productive lexical semantic relation between two concept types or classes that belong to incompatible or orthogonal types (type shift) This relation often has a directionality from a base sense to a derived sense Other terms used for this phenomenon are *regular polysemy* (Apresjan 1973) *sense extension* (Copestake 1995) and *transfers of meaning* (Nunberg 1996) The related concepts are lexicalized by the same word form in one language

Lexicalization patterns of these metonymic relations vary from one language to another Some languages may realize these regularities by the same word (which leads to polysemy), other languages by linguistic processes such as derivation and compounding

Metonymic relations between concepts in the ILI can thus be encoded independently of their realization in languages In practice this means that each

wordnet can represent its language-specific regular polysemic patterns within the ILI Classification is provided by a label to indicate from which language the metonymic cluster originates The metonymic relations can be identified by exploiting structural properties of any of the wordnets in the form of a class intersection of different senses of the lexicalized form

In order to distinguish types and instances of regular polysemy in WordNet1 5 we examined combinations of WordNet1 5 unique beginners There are 24 of these and each starts a unique branch in the WordNet hierarchy Examples are *artifact* and *substance* We started from the hypothesis that if their combinations subsume synsets that share the same word form this may reflect potentially regular semantic patterns at a very general level of description A similar approach was followed by (Buitelaar 1998), although we limited ourselves to combinations of two unique beginners, whereas Buitelaar investigated more than two

Our findings (Peters and Peters 1999) were that clustering on the basis of particular unique beginner combinations

1 regularly leads to odd clusters,
2 results in groupings that are not homogeneous in the sense that they do not display the same metonymic relation,
3 prevents the identification of subgroups that are semantically more homogeneous

In order to find these subgroups we identified nodes at a more specific level in the ontology whose combinations are shared by three or more words as hypernyms These words should occur in synsets that are hyponyms of these nodes at a distance of no more than 3 in terms of node traversal After manual verification we identified a number of fine-grained regular polysemic relations that are systematically encoded as sense distinctions of 105 words in WordNet A few examples

Under the unique beginner combination **artifact – substance** we found the relation *fabric/textile - fibre* (*cotton , alpaca fleece horsehair wool*)

Under the unique beginner combination **artifact – group** we found the relation *building - organization* (*academy body chamber room establishment school university club*)

It must be mentioned that some of these metonymic patterns are covered in a manually created table of 105 node pairs in WordNet1 5 (226 in WordNet1 6) that functions as the basis for the 'Relatives" search in WordNet All words with senses that are hyponymic to both nodes in a pair are grouped in the WordNet interface when similarity of meaning is queried However this grouping does not provide labels such as the ones above, nor does it guarantee that a cluster on the basis of one node pair is homogeneous

As a verification of the cross-linguistic validity of the regular polysemic patterns these language specific patterns can be projected from their source language onto the other EuroWordNet languages and it can be investigated whether they have corresponding lexicalization patterns

If the metonymic pattern occurs in several languages we have stronger evidence for the universality of the metonymic pattern

If there are no identical lexicalizations found in any other target language, or, in our case target language wordnet, there are three possibilities

1 the metonymic pattern is language specific and is not realised as a polysemous word in the target language For example, the Dutch *kantoor* is synonymous to the English *office* in the sense 'where professional or clerical duties are performed', but its sense distinctions cannot mirror the sytematic polysemic relation in English with 'a job in an organization or hierarchy'
2 The missing sense can in fact only be lexicalized by another word or compound or derivation related to the word with the potentially missing sense For example, the Dutch *vereniging* has the sense ("an association of people with similar interests") The English equivalent is *club* for which there is another sense in Wordnet ('a building occupied by a club') This is not a felicitous sense extension for the Dutch *vereniging*, because the favoured lexicalization is the compound *verenigingshuis* whose head denotes a building
3 The senses participating in the metonymic pattern are all valid senses of the same word in the target language, but one or more of them have not yet been captured in the wordnet For example, *embassy* has one sense in WordNet ('a building where ambassadors live or work) The Dutch translational equivalent *ambassade* has an additional sense denoting the people representing their country This sense can be projected to WordNet as a regular polysemy pattern that is also valid in English In fact LDOCE (Procter 1978) only lists the sense which is missing in WordNet

These metonymic sense groupings and their projections from the language in which they originate to other languages indicate a potential for enhancing the compatibility and consistency of wordnets (Peters et al , 1998) Verification will give an insight into the universality and productivity of these patterns Also, where languages display different

	clusters	words	word senses	synsets
Nouns	1703	1398	3205	2895
Verbs	2905	1799	5134	3839

Table 3 Statistics on Generalization clusters

lexicalization patterns, they can be used to derive semantic relations across wordnets, for instance a Location relation between the Dutch *vereniging* and *verenigingsgebouw*

3 2 Generalization

Clusters based on generalization consist of WordNet1 5 sense distinctions that are fine-grained enough to be grouped into a cluster with a more general meaning The fact that they are based on English lexicalization patterns is no methodological drawback because of the fact that the initial ILI merely consisted of WordNet senses

The clustering results in a reduction of ambiguity for polysemous words in WordNet and will indicate semantic relatedness between the senses of the synset members whose sense distinctions do not cover all clustered senses If necessary the original level of fine–grainedness can be restored by expanding the clusters into their constituent concepts

An incremental creation of larger clusters on the basis of a partial overlap between the existing clusters will enable us to create a layered status typology of ILIs and clusters involved and provide an interesting indications towards the standardization of word senses

In EuroWordNet the criterion of clusterable fine-grainedness has been operationalized by automatic means exploiting

- the internal hierarchical structure of Wordnet1 5, e g where two senses of a word share the same hypernym,
- many-to-one links between WordNet and other resources such as the Levin semantic verb classes (Levin 1993) (Dorr and Jones, 1996) and WordNet1 6
- cross-linguistic evidence many-to-one links between the ILI and the wordnets

A more detailed description of the various clustering methods can be found in (Peters and Peters 1999)

Table 3 gives an overview of the generalization clusters

3 3 Experimental results

To measure the effect of the ILI clusters we have automatically extended the sets of ILI-references for Dutch Italian and Spanish (as given in Table 1) with additional ILI cluster members that belong to the same cluster as any existing local concept For the nouns we see only a very small increase of about 1 to 1 5% For example, the total intersection for all 4 languages increased from 7736 (23,8%) to 8183 (25,2%) This is explained by the fact that the clusters only make up a small proportion of the total set of nouns

However, if we look at the verbs we see a doubling of the total intersection from 1632 (21,9%) to 3051 (40,9%) Since relatively many verbal clusters have been added and since the number of verbs synsets is much lower than the noun selection such a strong effect makes sense We therefore can expect a much bigger effect of the verbal clusters in Word-Sense-Disambiguation and Information-Retrieval than for the nouns

4 The ILI as the superset of word meanings

As explained in the introduction, the ILI should be the superset of all the concepts occurring in the different wordnets so that we can establish relations between minimal pairs of synsets Initially, the index was based on the synsets that occur in WordNet1 5 However, in the other wordnets there may be concepts that do not occur or cannot be found in WordNet1 5 These concepts are, for the time being manually linked by means of complex equivalence relations to other closely related, concepts in the ILI For example, the Dutch concept *klunen* does not occur in WordNet1 5 but can be related by so-called complex equivalence relations to other concepts

klunen = {to walk on skates over land from one frozen water to another frozen water}
EQ_HAS_HYPERONYM *walk v*
EQ_INVOLVED *skate n*
EQ_IS_SUBEVENT *skate v*

Such synsets in the local wordnets which are not linked by an EQ_(NEAR)_SYNONYM relation to the ILI are potential candidates for new ILI-records The general procedure to further select ILI-candidates selects proposed concepts that occur in at least two languages and do not overlap with current concepts in WordNet1 5

Obviously we have to consider the relevance of these missing concepts for a universal list of sense-distinctions So far, we have carried out two different evaluations of potential sources of ILI records

- we inspected two sets of Dutch verbs that did not receive any translation to English using bilingual dictionaries,
- we compared two sets of proposed ILIs based on the German wordnet and the Italian wordnet with the Dutch wordnet to measure potential overlap

4.1 Evaluation of verbal Dutch mismatches

We have looked at two sets of Dutch verbs without translation

- 32 static verbs (hyponyms at 3 levels below *zijn* (to be))
- 41 dynamic verbs (hyponyms at 3 levels below *gebeuren* (to happen))

These verbs could either not be found in the bilingual dictionaries or their phrasal translation could not be matched to WordNet1 5 Some of the synsets could still be matched with some effort (3 static verbs and 5 dynamic verbs) The remaining unmatched concepts could be classified as follows

Matches to different Part of Speech· verbs that could be matched to an adjective or noun that has the same meaning (15 static and 5 dynamic verbs)

Exhaustive Links: verbs whose meaning is fully captured by several links to multiple ILI-records (6 static and 21 dynamic verbs)

Incomplete Links· verbs that can only be linked to a hyperonym ILI records that classifies it (4 static and 10 dynamic links)

Unresolved Links cases that cannot even be linked to a hyperonym ILI record (4 static verbs and 0 dynamic)

The first category contains part of speech mismatches For instance, for the static verb *aanstaan* (be ajar) there is no phrasal entry *be ajar* in WN1 5, but there is the adjective *ajar* which means open' Similarly the verb *bankdrukken* is translated as *benchpress* (without a space), but WN1 5 has the noun *bench press* which has the same meaning a weightlifting exercise In EuroWordNet we have decided that the ILI is part-of-speech neutral in the sense that words with a different part of speech can still be linked to each other Therefore EQ_NEAR_SYNONYM relations have been assigned to the adjective *ajar* and to the noun *bench press* It is thus not necessary to extend the ILI for concepts that match in meaning but have a different part of speech Strictly speaking, this would also imply that current ILI-records which are synonymous but have a different part of speech in English could be merged or grouped by composite ILIs as well just as the generalizations that we have discussed There is no need to have two concepts for *departure* and *depart* in the ILI, since both are conceptually equal and the realization in a language can be either as a verb or a noun, or by both (as in English)

The second category of unmatched verbs often follows a regular pattern, where the verb has a compound structure and its meaning is compositionally derivable from that structure, e g

doodvechten (fight to the death)
EQ_HAS_HYPER *fight* & EQ_CAUSES *death*

draadtrekken (produce a wire by pulling)
EQ_HAS_HYPER *produce/make* &
EQ_HAS_HYPER *pull* &
EQ_INVOLVED *wire*

The verb *doodvechten* means 'fight to the death which is not in WN1 5 Internally the hyperonym is *vechten* (fight) and there is a cause relation with *dood* (death) Both are also assigned as equivalents The verb *draadtrekken* means 'to make a wire by pulling' and is linked to the hyperonyms *pull* and *make/produce*, as well as to the result *wire* Typically, we see here that the meaning of these verbs is exhaustively covered by the multiple equivalent links Furthermore, it is possible to derive many more of these meanings productively and generate the corresponding verb compound in Dutch In general, if a synset has two hyperonyms or a hyperonym and another relation (CAUSE, INVOLVED MANNER, RESULT) there is often no need for a new ILI concept Just as with the cross-part-of-speech matches the above strategy would imply that current ILI-records that can be linked and predicted in the same way should be removed from the standardized list

The remaining cases are unsatisfying matches (18 in total, or 24%) These are all characterized by having assigned only one hyperonym or several near_synonyms or a combination of these and are therefore genuine candidates for new ILI concepts

For most unmatched verbs, it is thus not really necessary to extend the ILI Moreover we could apply the same analysis to the WordNet1 5 based ILI and further reduce it However, it is still necessary to know that the meaning is exhaustively captured by the equivalence relations and can uniquely be derived from these links Only in that case we can establish equivalence relations across languages by combinations of links A Dutch synset that is exhaustively linked by a hypernym and cause relation to the ILI would match an Italian concept only if it is linked exhaustively by the same equivalence relations and there is no other Italian synset linked in the same way (and vice versa) Unfortunately exhaustiveness has to be encoded manually This process

Disambiguation Strategy	Clustered synsets	Reductions on polysemous terms
Manual	10054/78902 (13%)	-
First Sense	10420/93240 (11%)	-
AR	24526/149632 (16%)	11936/29403 (41%)
No disambiguation	68515/387469 (18%)	49074/65737 (75%)

Table 4 Effects of the ILI clusters on the IR-SEMCOR text collection

can be helped by looking at the morpho-syntactic markedness (e g regular compound structures), regular lexicalization patterns and corpus frequency

4 2 Cross-linguistic overlap of mismatches

To get an idea of the cross-linguistic overlap of unmatched synsets such as the above, we have inspected a sample of the Italian and German mismatches to see if they could potentially overlap with Dutch synsets The Italian and German synsets have been selected because they had no straightforward mapping with the ILI after manual checking Comparison with a random sample of 36 German noun synsets showed that 50% of the nouns (18) have an equivalent in Dutch For a sample of 59 Italian noun synsets there is at least an overlap of 30% (20) with Dutch Examples are Arbeitszeitverkurzung (DE) = arbeidstijdverkorting (NL) = (reduction of working hours) and Baita (IT) = berghut (NL) = (cabin in the mountain)

If we quantify these results for the total Dutch wordnet, where about 6,000 Dutch synsets can not be translated, this would imply that at least 30% (2,000 synsets) represent new concepts that overlap with German or Italian, and therefore should be added to the ILI, although we feel that a native English speaker should verify the absence of the concept in English and in WordNet1 5

For the ILI-verbs it is much more difficult to give any numbers For German only 10 ILI-verbs are proposed It is not possible to draw any conclusions from such a small set The number of Italian ILI-verbs is about 70 and it is clear that the overlap with Dutch is very low This is due to the fact that many proposed verbs (50%) are multi-words in Dutch, e g *abbuiarsi* (get serious) *infiacchire*(make lazy) Just as the Dutch verbs in the previous subsection, many of these can be assigned with an EQ_HYPERONYM and EQ_CAUSES to WN1 5 and therefore do not have to be added as a new ILI concept The remaining cases are too difficult to judge, and more information is needed to understand the intended concept

For verbs we thus expect that the number of new ILIs will be relatively low First of all, there not many synsets that do not have translations (compared to nouns) and secondly, unmatched verbal synsets often can be linked somehow exhaustively

5 Using the ILI as a standardized meanings in NLP

The ILI provides a language-neutral conceptual map for -especially multilingual- NLP applications For instance, a multilingual text collection can be indexed in terms of the ILI records, obtaining a uniform representation for documents, regardless of their particular languages Such a representation can be used to perform language-independent Text Retrieval This approach differs substantially from the mainstream Cross-Language Text Retrieval strategy, namely translating the query into the target languages, using bilingual dictionaries, bilingual corpora or Machine Translation systems Some advantages of indexing with ILI records are

- It distinguishes different senses of a word, in any language,
- It conflates synonym terms within and across languages,
- It scales up to more than two languages better than query translation approaches,
- Terms can be related not only by identity, but on the basis of more sophisticated relations (Cross Part-of-Speech relations, hyponymy, meronymy, etc) This allows for more sophisticated, and language-independent weighting and retrieval

In spite of its appeal, this approach is challenging because

- It demands accurate word-sense disambiguation to restrict the possible ILI records for a given term,
- It should exploit EWN conceptual relations to associate
  - Strongly related terms that differ in POS (through XPOS relations) For instance, a standard IR system does not distinguish between the verbal and nominal form of *design* which can be an advantage in many retrieval situations But in EWN they are mapped to different synsets in different hierarchies Only XPOS relations (absent in WordNet) permit to establish the appropriate connection,

Monolingual Experiments

	Text	Manual WSD	First sense	AR	No WSD	Manual queries
Wn1.5	31 7	35 7(+12 4%)	31 7(=)	32 2(+1 4%)	30 2(-4 8%)	33 4(+5 1%)
ILI	=	35 4(+11 5%)	31 7(=)	32 1(+1 2%)	30 2(-4 8%)	33 2 (+4 6%)

Cross-Language (Spanish to English) experiments

	Dict. expansion	Manual WSD	AR	No WSD	Manual queries
EWN	23 9	32 1(+34 5%)	21 1(-11 9%)	20 7(-13 2%)	31 1(+30 1%)
ILI	=	32 0(+33 9%)	20 7(-13 2%)	20 5(-14 2%)	31 1(+30 1%)

Table 5 Information Retrieval experiments with different WSD strategies

- Strongly related meanings of a word that usually discriminate the same context (through ILI clusterings)

- It has a higher computational cost (at indexing time) to map documents into the ILI

We have conducted some experiments to test a) how different WSD strategies affect precision/recall figures, and b) how ILI clustering may affect indexing and retrieval performance We have used a variation on the IR-SEMCOR test collection described in (Gonzalo et al , 1998) This test collection, adapted from Semcor, is small for current IR standards (3Mb excluding all tags, slightly bigger than the standard TIME collection), but is fully semantically tagged This feature permits comparing the performance of manual versus automatic sense disambiguation / sense filtering The set of queries is available and hand-tagged in English and Spanish, permitting monolingual and Cross-Language (Spanish to English) retrieval

The results are shown for a number of different indexations of the IR-SEMCOR collection, with and without using the actual ILI clusters There are three *full disambiguation* strategies in which every noun term is represented as a single synset The rest are *sense filtering* strategies that return the list of more likely synsets for every noun term Words other than nouns are left unchanged

The disambiguation strategies are

Manual returns synset assigned by IR-SEMCOR tags

First sense Returns First sense in Wordnet 1 5 (not applicable on Spanish queries),

AR (Agirre-Rigau) An implementation of the Agirre-Rigau WSD algorithm (Agirre and Rigau, 1996), that has the advantages of a) being unsupervised and b) being applicable on any language, provided there is a WordNet for it This algorithm gives a weighting for the candidate senses, rather than just picking one of them and discarding the rest In the experiment we take all the senses with maximal weight Its WSD performance is lower than the First Sense heuristic, especially disambiguating queries, as the disambiguation context is much smaller,

No WSD A noun term is represented with all its possible synsets,

Manual queries Combines the *No WSD* strategy for documents and the *Manual* strategy for queries This is a plausible combination of efficient document indexing (no disambiguation is required) with interactive retrieval (user-assisted disambiguation)

Table 4 shows how the ILI clusterings reduce ambiguity in the representation of the documents for each of the indexing strategies The first column in the table shows the number of clustered occurrences of noun synsets against the total number of noun synsets The second column shows the number of reductions performed on ambiguous terms (that is on terms that are not fully disambiguated and are thus represented as a list of synsets) One reduction means, e g that a word represented as n different synsets is now represented as $n-1$ different synsets

The number of clustered synsets is quite high, given the small size of ILI noun clusters In particular the ambiguity reduction is very promising with 49074 reductions in 65737 polysemous terms in the collection The reason is that clusters are mostly applied on highly polysemous words, which are in turn the most frequently used

The results of the monolingual and cross-language IR experiments can be seen in Table 5 The results without clusterings are in the first row and with clustering in the second row The figures represent the average precision at ten fixed recall points between 10 and 100 We have used the INQUERY system (Callan et al , 1992) to perform the experiments The results suggest

- There is a potential improvement over standard INQUERY runs as shown by the results on

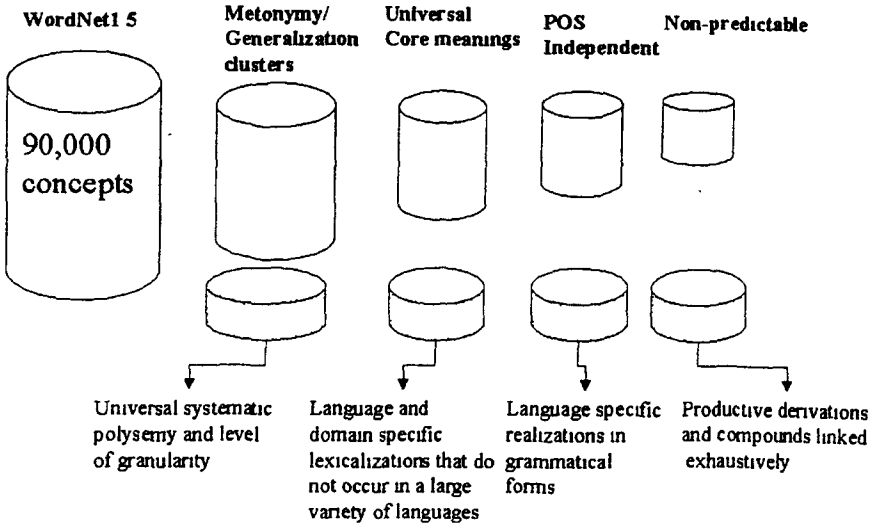

Figure 1 From WordNet to ILI

the manually disambiguated collections The Cross-Language track is especially promising, with a gain of 34 5% over the standard technique (translation of the query using POS tagging and bilingual dictionary expansion)

- Although the Agirre-Rigau algorithm performs much worse than the First Sense heuristic in terms of WSD accuracy, it gives slightly better results for IR, as it just filters the most unlikely senses This is experimental evidence in favor of evaluating WSD algorithms within concrete tasks, in addition to general-purpose evaluations such as the SENSEVAL one
- The last column ('manual queries") corresponds to expansion to all synsets in the documents (no disambiguation) and manual disambiguation of the query This method improves Cross-Language Retrieval by 30% (comparable to full manual indexing), and degrades only 7% from monolingual to bilingual retrieval (standard degradation is 30-60%) This suggests that EWN can be very useful in interactive retrieval settings (where the user is guided through a disambiguation process) even if the database has not been disambiguated at all
- The results using the ILI clusters are similar or slightly worse than without clustering A possible reason is that the ILI clusters and the clusters needed for IR do not exactly match It would be probably beneficial to further distinguish types of clustering according to their ability to identify co-occurring senses of a word, in a similar vein to Buitelaar's white and black dot operators (Buitelaar, 1998) These operators distinguish related senses that tend to co-occur simultaneously (such as *book* as *written work* or *physical object*) and related senses that occur in different contexts (such as *gate* as *movable barrier* or *computer circuit*) Obviously, the first ones are optimal candidates for clustering in Information Retrieval applications

  A more refined typology of ILI clusterings in general, seems required to use different clustering types for different tasks

6 Conclusions

We described the building of a universal list of meanings in EuroWordNet the so-called Inter-Lingual-Index (ILI), for which Wordnet1 5 was taken as a starting point The ILI should provide an efficient mapping between concepts across languages For that purpose it should have a certain granularity and completeness with respect to the sense-differentiation found in the wordnets for different languages

We provided empirical evidence for a more univer-

sal and efficient level of sense-differentiation based on structural properties of the wordnets and their multilingual mapping and alignment This has lead to a typology of sense-distinctions, where the status of ILI-records can be differentiated along the following lines

- Universality In how many languages does the concept occur? How universal is polysemy?
- Usage how frequent is a concept used across languages?
- Productivity how easily can similar or related concepts be derived as new concepts?
- Exhaustiveness how complete and unique can a concept be linked to other concepts?
- Dependency can concepts be related by (semi-)productive sense extension and how universal are these extensions?
- Morpho-syntactic markedness do words have a systematic morpho-syntactic structure across languages?
- Ontological status to which degree can concepts be distinguished in a minimally overlapping way?

These criteria can be used to create a minimalized and efficient list of sense-distinctions Not all missing sense-distinctions from other wordnets should be added to WordNet1 5, where productivity and predictability can be captured via exhaustive complex mapping relations Furthermore, other sense-distinctions could be generalized or grouped Figure 1 gives an overview how these criteria can be used to reduce the initial fund of concepts, as discussed in this paper

The restructuring of ILI and the development of a universal core list of word meanings is useful to

- more efficiently map wordnets across languages,
- more efficiently apply WSD and Cross-Language IR (XL-IR),
- apply the same WSD/XL-IR across languages,
- verify WSD/XL-IR techniques across languages

Some experimental results demonstrating this have been reported, but a lot of work still needs to be done We hope that the ILI could be used in a new round of SENSEVAL/ROMANSEVAL to demonstrate the capacity to compare and apply WSD technologies cross-linguistically We think also that the ILI is an interesting resource to experiment semantically-oriented approaches to Multilingual Information access tasks such as Cross-Language Text Retrieval in the reported experiment

References

Eneko Agirre and German Rigau 1996 Word Sense Disambiguation using Conceptual density In *Proceedings of COLING'96*

J Apresjan 1973 Regular polysemy *Linguistics*, 142

P Buitelaar 1998 *CoreLex Systematic Polysemy and Semantic Underspecification* Ph D thesis, Department of Computer Science, Brandeis University, Boston

J Callan, B Croft, and S Harding 1992 The INQUERY retrieval system In *Proceedings of the 3'd Int Conference on Database and Expert Systems applications*

A Copestake 1995 Representing lexical polysemy In *Proceedings of AAAI Stanford Spring Symposium*

B Dorr and D Jones 1996 Role of Word Sense Disambiguation in lexical acquisition Predicting semantics from syntactic cues In *Proceedings of the Int Conference on Computational Linguistics*

C Fellbaum 1998 A semantic network of English the mother of all wordnets *Computers and the Humanities, Special Issue on EuroWordNet*

J Gonzalo, F Verdejo, I Chugur, and J Cigarran 1998 Indexing with Wordnet synsets can improve Text Retrieval In *COLING/ACL'98 Workshop on Usage of WordNet in Natural Language Processing Systems*

B Levin 1993 *English Verb Classes and Alternations* Univ of Chicago Press

G Numberg 1996 Transfers of meaning In J Pustejovsky and B Boguraev, editors, *Lexical Semantics The problem of polysemy* Clarendon Press

W Peters and I Peters 1999 Restructuring the InterLingual Index Technical report EWN Deliverable 2d004

W Peters, I Peters, and P Vossen 1998 Automatic Sense Clustering in EuroWordNet In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 98)*

P (ed) Procter 1978 *Longman Dictionary of Contemporary English* Longman Group

Y Wilks 1998 Is Word Sense Disambiguation just one more NLP task? *Computer and the Humanities, Special Issue on Senseval*