Extending the Inter-Lingual-Index with new concepts

Version 2, Final

2-8-1999

Contributors:

Piek Vossen, Laura Bloksma, University of Amsterdam

Wim Peters, University of Sheffield

Claudia Kunze, Andreas Wagner, University of Tübingen

Karel Pala, University of Masaryk

Kadri Vider, University of Tartu

Francesca Bertagna, Istituto di Linguistica Computazionale - CNR, Pisa

Deliverable 2D010
EuroWordNet, LE2-4003

Identification number	LE-4003-2D010
Type	Document and Lingware
Title	Extending the Inter-Lingual-Index with new concepts
Status	Final
Deliverable	2D010
Work Package	WP11
Task	T11.2
Period covered	December 1998 – July 1999
Date	2-8-1999
Version	2
Number of pages	33
Authors	• Piek Vossen, Laura Bloskma, University of Amsterdam • Wim Peters, University of Sheffield • Claudia Kunze, Andreas Wagner, University of Tübingen • Karel Pala, University of Masarky • Kadri Vider, University of Tartu • Francesca Bertagna, Istituto di Linguistica Computazionale - CNR, Pisa
WP/Task responsible	SHE
Project contact point	Piek Vossen University of Amsterdam Spuistraat 134 1012 VB Amsterdam The Netherlands tel. +31 20 525 4669 fax. +31 20 525 4429 e-mail: Piek.Vossen@hum.uva.nl
EC project officer	Ray Hudson
Status	Public
Actual distribution	Project Consortium, the EuroWordNet User Group, the Ad Hoc Ansii Committee for Ontology standards, the world via http://www.let.uva.nl/~ewn.
Suplementary notes	n.a.
Key words	Linguistic Resources, Ontologies, Multilingual Wordnets, Language Engineering

Abstract	This deliverable describes an evaluation of the possibilities to extend the Inter-Lingual-Index with new concepts. Two different extensions have been considered: - extension with domain terminology - extension with concepts from other languages that have not been found in WordNet1.5. Only the first extension has been implemented in the EuroWordNet database for the domain of computing.
Status of the abstract	Complete
Received on	
Recipient's catalogue number	

Executive Summary

In this deliverable we describe the extension of the Inter-Lingual-Index (ILI) with new concepts that are not present in WordNet1.5. There are three different kinds of extensions to be considered:

1. concepts that are specific for a domain;
2. concepts that should be in WordNet1.5 as a generic English resource but are missing;
3. missing concepts that are specific for a non-English language;

To see how EuroWordNet, and consequently the ILI, can be extended for a domain, we have adapted the ILI for the domain of computing. For this purpose, the computer terminology that was already present in the ILI (based on WordNet1.5) was labeled accordingly, and new concepts from domain sources were added. The extension with computer terminology is available in the EuroWordNet database and can be accessed via the domain labeling. Other missing concepts in WordNet1.5 are detected by trying to link synsets in the local language to the appropriate equivalences in the ILI.

We describe the proposals for new concepts based on the Dutch, German, Italian, Estonian and Czech wordnets, and an evaluation of these proposals. Finally, we describe a proposal to develop a condensed and universal index of meaning. This proposal was presented in a paper at the Siglex workshop of ACL-99 in Maryland.

Table of Contents

1. Introduction

EuroWordNet is a multilingual database with generic wordnets for 8 different languages: English, Dutch, Spanish, Italian, German, French, Czech and Estonian. Each of these wordnets is structured along the lines of the Princeton wordnet (Fellbaum 1998) in terms of sets of synonyms or synsets between which basic semantic relations are expressed. In addition, most of the synsets in each language are connected to a so-called Inter-Lingual-Index. Synsets connected to the same index records can be seen as conceptual equivalences across the languages.

The index is initially based on the synsets or concepts in WordNet1.5 but is adapted to provide a more efficient mapping across the wordnets. In this deliverable we describe the extension of the Inter-Lingual-Index (ILI) with new concepts that are not present in WordNet1.5. There are three different kinds of extensions to be considered:

4. concepts that are specific for a domain;
5. concepts that should be in WordNet1.5 as a generic English resource but are missing;
6. missing concepts that are specific for a non-English language;

To see how EuroWordNet, and consequently the ILI, can be extended for a domain, we have adapted the ILI for the domain of computing. For this purpose, the computer terminology that was already present in the ILI (based on WordNet1.5) was labeled accordingly, and new concepts from domain sources were added. The extension with computer terminology is available in the EuroWordNet database and can be accessed via the domain labeling. This work is described in section 2.

Other missing concepts in WordNet1.5 (2 and 3) are detected by trying to link synsets in the local language to the appropriate equivalences in the ILI. This work is reported in section 3 and 4, where section 3 describes the proposals for new concepts based on the Dutch, German, Italian, Estonian and Czech wordnets, and section 4 contains a comparison and evaluation. Finally, we describe a proposal to develop a condensed and universal index of meaning. This proposal was presented in a paper at the Siglex workshop of ACL-99 in Maryland.

2. Extending the Inter-Lingual-Index for the domain of computing

Terminology specific to the domain of computing has been incorporated semi-automatically into the EuroWordNet general language resource involving various web-based terminological resources and manual evaluation. Because of time constraints we have aimed at quality and representativeness, not at exhaustiveness. The set forms a representative sample indicating the feasibility of incorporating domain specific terminology into the EuroWordNet general language resource. The selection process has concentrated on English terminology. The implementation of the domain terminology in other languages has relied as much as possible on resources containing translational equivalences for the selected computer terms.

The following freely available English resources have been used in the selection process:

FOLDOC Free On-line Dictionary of Computing, Editor Denis Howe
Around 6000 entries with definitions and subdomain information
http://wombat.doc.ic.ac.uk/foldoc/index.html

DATA Direct glossary
http://data-direct.com/glossary.htm
around 650 entries with definitions

Dartek glossary
http://www.dartek.com/glossary/glossary.cfm
around 1000 entries with definitions

Netglos glossary
http://wwli/com/translation/netglos/netglos.html
around 110 entries with definitions

The following criteria for selection have been applied in decreasing order of importance:

1. The presence of the word form in WordNet1.5.
   In order to maximise the conceptual coverage of the EuroWordNet vocabulary all word forms present in WordNet1.5 were selected. We first investigated the senses of these terms, and found that 107 occur with matching domain-specific senses, e.g.:

   Local area network 1: "a network connecting computers and word processors and other electronic office equipment to create an inter-office system"

   Glosses, where missing in WordNet1.5, were added on the basis of the available online computing dictionaries.

   A further 168 were found in WordNet1.5 as word form but without a computing sense. 18 of these senses are hyponyms of existing senses in WordNet1.5. These latter senses are a broader, more general semantic characterisation of the computer terms, but may also cover entities outside the domain of computing.
   Some examples:

   Integrity 1: "an unreduced or unbroken completeness or totality." is the hypernym of the newly added computing sense described by the gloss "The quality of being uncorrupted".

   Packet 1: "a collection of things wrapped or boxed together" has as hyponym the new computing sense: "A sequence of data, with associated control information, that is switched and transmitted as a whole; refers mainly to the field structure and format defined with the CCITT X.25 recommendation. In data transfer, information is broken into packets, which then travel independently through the Net. "

   The remaining 150 new terminological senses were added to existing word forms in EuroWordNet. Some word forms received 2 new computing senses:

   pointer 4: "An address, from the point of view of a programming language. A pointer may be typed, with its type indicating the type of object to which it points."
   pointer 5: "A link to related resources inserted into a Web page."

2. Frequency information in several text resources:

Another 169 have been selected that do not occur in WordNet1.5. Frequency information from the following resources has been used:

- Ami-Pro manuals
- British National Corpus
- Unix manuals

The final set of selected computer terms consists of 439 terms (393 nouns, 32 verbs and 14 adjectives), listed in Appendix II.

All these terms have received a domain label. For the majority of terms this label is "Computing". Six subdomains of computing (taken from Foldoc) appeared to cover more than one term in this set. These have been added as domain labels for 85 terms:

World Wide Web	8
Networking	8
Storage	7
Programming	30
Operating System	19
Hardware	13

3. Extending the Inter-Lingual-Index with concepts from other languages

The Inter-Lingual-Index (ILI) in EuroWordNet should connect synsets across all the languages. In practice this means that the ILI should be the superset of all the concepts occurring in different wordnets (at least two). Initially, the index was based on the synsets that occur in WordNet1.5. However, there may be concepts in the other wordnets that do not occur or cannot be found in WordNet1.5. These concepts are, for the time being, 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"
EQ_INVOLVED "skate"
EQ_IS_SUBEVENT_OF "speed skating"

Such synsets in the local wordnets, which are not linked by a eq_(near)_synonym relation to the ILI are potential candidates for new ILI-records. The potential procedure to further select ILI-candidates is as follows:

1) each site selects a set of 'unlinked' synsets and derives a potential new ILI-record, specified as described below.
2) AMS/SHE collect these proposals compare them automatically
3) Likely candidates are checked manually
4) ILI-records proposed by two languages will be accepted
5) All partners check this list and try to find matching synsets
6) SHE verifies whether the proposed ILI-records do not occur in WordNet1.5
7) new records are added and the each site updates the references from their wordnet to these records if they have a matching word or expression.

Proposed ILI-records that are not matched across languages are kept in a separate fund for future extensions.

To be able to match and compare the proposed ILI records they have to be specified in the following format:

```
0 POSSIBLE_ILI_RECORD
  1 PART_OF_SPEECH "v"
  1 NEW_ILI_ID
  1 NEW_ILI_SOURCE "Dutch"
  1 GLOSS "to walk on skates over land to go from one frozen water to
another frozen water"
  1 VARIANTS
    2 LITERAL "walk on skates"
      3 SENSE 1
  1 SOURCE_VARIANTS
    2 LITERAL "klunen"
      3 SENSE 1
  1 EQ_LINKS
    2 EQ_RELATION "eq_has_hyperonym"
      3 TARGET_ILI
        4 PART_OF_SPEECH "v"
        4 FILE_OFFSET 254093
        # hyperonym is "walk"
    2 EQ_RELATION "eq_involved"
       3 TARGET_ILI
        4 PART_OF_SPEECH "n"
        4 FILE_OFFSET 254093
        # instrument is "skate"
    2 EQ_RELATION "eq_is_subevent"
       3 TARGET_ILI
        4 PART_OF_SPEECH "v"
```

```
      4 FILE_OFFSET 254093
      # is subevent of "speed skating"
1 TOP_CONCEPTS
     2 TOP_CONCEPT "Dynamic"
     2 TOP_CONCEPT "Location"
     2 TOP_CONCEPT "Usage"
     2 TOP_CONCEPT "Physical"
```

The gloss can be based on a phrase from the bilingual dictionary and should be in English. Also the synset variants have to be English phrases. These may be derived from the gloss (e.g. the genus and a simple modifier). In addition to the ILI-variants there should be source-variants. The source-variants are directly copied from the source wordnet. These will be used to match new ILIs from closely related language (e.g. Spanish-Italian, Dutch-German). By matching the left part of the variants we can derive an equality score. The EQ_LINKS are used to cluster the potential ILIs semantically. These can be automatically derived from the current language-internal relations. A minimal requirements is that the new ILI has an EQ_HYPERONYM, other relations are optional. The Top-Concepts (TCs) can be derived from the language-specific wordnet by inheriting them via the hyperonym relations from the Base Concepts. The TCs will also be used for semantically clustering potential ILIs. Most of the above information can be derived automatically from the current wordnets.

In the next subsections, we will describe the results of a first evaluation of new ILI-proposals extracted from the Dutch, German, Italian and Estonian wordnets. In Appendix I, we also give a list of Czech nouns that could not be translated.

3.1. Selection of the Dutch candidates for new ILI-records.

Most synsets in the Dutch wordnet are translated by automatically looking up the synset members in a bilingual dictionary (Martin and Tops 1989) and by looking up the translations in WordNet1.5. The senses of the translations are selected by a series of different heuristics. This automatic mapping procedure is implemented in a local database, in which the lexical resources are loaded (see Vossen, Bloksma and Boersma 1999, for more details). About 22% of the noun translations (8,596 equivalences) and 55% of the verb translations (6,155 equivalences) was checked or assigned manually. Manually assigned translations also include complex equivalence relations such as the ones described above for the example “klunen”. In total, 6,070 nominal synsets (16%) and 1,133 verbal synsets (12%) did not receive a reference to a ILI-record.

For the Dutch, we considered 3 classes of possible candidates for new concepts:

1. Dutch synsets that have not received a translation after the automatic mapping procedure.
2. Non-equivalent Dutch synsets that are linked to the same ILI-record via a EQ_NEAR_SYNONYM relation.
3. Dutch synsets that have been linked manually via a complex equivalence relation: i.e. not via an EQ_SYNONYM or EQ_NEAR_SYNONYM relation.

Two samples of the first group have been inspected manually to see if they represent a good source for potential new concepts. The inspected synsets without an equivalent relation were selected from the hyponyms below the two top verbs in the Dutch Wordnet, 'zijn_7' *(be)* and 'gebeuren_2' *(happen)*.

verb tops	zijn_7 *(be)*	gebeuren_2 *(happen)*
Number of hyponyms 3 levels down	305	1311
no equivalent	32	166

Below the top 'zijn_7' *(be)* the total number of hyponyms three levels down were 305 of which 32 hyponyms had not received a translation.

For instance for the verb 'aanstaan_1' *(be ajar)* there is no entry 'be ajar' in WN1.5. There is however the adjective 'ajar' which means 'open'. Therefore an EQ_NEAR_SYN relation to the adjective 'ajar' has been assigned. Another example of this kind is:

Dutch verb	translation	equivalent	WN1.5
bijeenzijn_1	be together		
	be gathered	EQ_NEAR_SYN	together_6 (adverb)

We found 15 synsets with the translation pattern 'be Adj/Adv' without an equivalent entry in WN1.5 with the same POS. This is due to different factors. For the improvement of WN1.5 it may be interesting to look at these cases (by a native speaker of English). Whether a new concept has to be added to the ILI for these cases is a different matter. We defined across POS synonymy as an internal relation (internally an XPOS_NEAR_SYN relation with an adjective or adverb is often assigned for the examples described above). So if an across POS equivalent_near_synonym can be assigned there is no need for adding a new concept to the ILI.

No equivalent was found for manner compound verbs (V+A) like blindvaren_1 *(trust blindly)*, because no such verb exists in WN1.5. Internally in Dutch there is a hyperonym 'vertrouwen' *(trust)* and a manner_relation with the adjective blindelings *(blindly)*. Therefore both an equivalent_has_hyperonym with 'trust' and an equivalent_manner relation with 'blindly' has been assigned.

No equivalent was found for compound verbs (V+P) like bijslapen_1 *(catch up some sleep)*. The verb has internally two hyperonyms: 'slapen' *(to sleep)* and 'inhalen' *(to catch up)*. These hyperonyms both have an equivalent WN1.5 the first to 'sleep' and the second to 'catch up'. Therefore both synsets are linked as an equivalent_hyperonym to 'bijslapen_1'.

Another example:

Dutch		translation	WN1.5
doorslapen_1		(sleep on/sleep through)	
internally 2 hyperonyms:			
slapen_1	(sleep)	EQ_HAS_HYPERONYM	sleep
doorgaan_1	(continue)	EQ_HAS_HYPERONYM	continue

If a synset has two hyperonyms or a hyperonym and a manner relation there is no need for a new ILI concept.

No equivalents were assigned to verbs for which somehow the translation given in the bilingual dictionary did not match with a concept in WN1.5, but with some effort an equivalent could be found. For example slapen_5 (lit.: be asleep/not functioning) has no translation at all in the bilingual. In WN1.5. the concepts 'lie dormant, be inactive' and 'run idle' come close. So both equivalents have been assigned as near synonyms.

Other example:

Dutch	equivalent	WN1.5
thuisblijven_1 (stay at home, stay in)	EQ_N_S	stay/ remain behind

Synsets with unsatisfying matches are all characterized by having assigned only one hyperonym or two near_synonyms or a combination of this and therefore are possible candidates for new ILI concepts.

Overview

EQ_NEAR_SYN with adjective	15
EQ_HAS_HYPER (2)	3
EQ_HAS_HYPER and EQ_MANNER = adjective	3
EQ_NEAR_SYN	3
EQ_HAS_HYPERONYM	4
not solved	4

Below the verb top 'gebeuren_2' *(happen)* the total number of hyponyms were 1311 of which 166 had no equivalent. We restricted the analysis to 41.

OVERVIEW	
pattern: (be adjective) EQ_N_S adjective	3
EQ_N_S verb resolved manually	5
EQ_N_S noun resolved manually	2
EQ_HAS_HYPER (2)	3
EQ_HAS_HYPER and EQ_NEAR_S	1
EQ_HAS_HYPER and CAUSES	10
EQ_HAS_HYPER and INVOLVED	4
EQ_HAS_HYPER and MANNER	2
EQ_HAS_HYPER and RESULT	1
EQ_HAS_HYPER	10
Total	41

Again many equivalents could be resolved. This is however time-consuming but it cannot be done automatically because a missing equivalent can be due to so many factors.

There were two cases were a noun equivalent could be assigned. The verb bankdrukken_1 is translated as 'benchpress'. In WN1.5 there is a noun (with a space in between) 'bench press' which has the same meaning. The verb brandstichten_1 is translated as 'commit arson'. In WN1.5 there is the synset {arson, fire-raising, incendiarism} which is equivalent. This noun has an equivalent relation with the Dutch noun: brandstichting (arson) and therefore an XPOS-relation between the Dutch verb and noun should be there also.

The next cases are examples that involve a hyperonym equivalent together with another equivalent (CAUSE, INVOLVED, MANNER, RESULT).

EQ_HAS_HYPER and EQ_CAUSE relation
For example 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.

EQ_HAS_HYPER and EQ_INVOLVED relations
'omdraaien' (radical change of opinion)

EQ_HAS_HYPER	change
EQ_INVOLVED	opinion

EQ_HAS_HYPER and EQ_MANNER
'rouwdouwen' (act in a rough way)

EQ_HAS_HYPER	act
EQ_MANNER	roughly

EQ_HAS_HYPER and EQ_RESULT
'draadtrekken' (no translation, means lit.: produce into a wire by pulling)

EQ_HAS_HYPER	produce/make
EQ_INVOLVED	wire

The 10 synsets with just one equivalent_hyperonym relation are the only possible new ILI concepts.

A second source of possible new ILIs, are that multiple Dutch synsets are linked to the same ILI-record. These concepts only represent a weak source. They may indicate the following:

- One or both are wrongly translated (especially if the Dutch synsets are very different).
- The Dutch synsets should be merged or they should be linked by a NEAR_SYNONYM relation.
- The Dutch synsets are sufficiently different and correctly translated.

If there is sufficient reason to keep the Dutch synsets separate as different concepts and the translations are correct, then it is also arguable that the ILI should be differentiated. However, this situation only makes sense when the differences are not too extreme. This may be the case when the Dutch senses are more specific than the ILI-records, or when they represent a level in between the ILI-records and its hyperonym. Still, the difference

should be rather subtle, because they would otherwise motivate a EQ_HYPERONYM or EQ_HYPONYM relation.

The final group is the most likely source for new concepts. In this case no equivalences have been found in the ILI, even after manual inspection. Only if the synset has one or more translations in the bilingual dictionary it has been proposed as a new record (there are also cases that could not be matched with the bilingual dictionary). This translation (often a phrase) could not be matched with a WordNet1.5 entry. The gloss and the English variant are extracted from the translations in the bilingual dictionary. The Top-Concepts are inherited from the Dutch hierarchy. Currently, only this set has been used to generate the potential ILI-records in the above experiment.

3.3. Selection of the German candidates for new ILI-records.

After building subset 1, we selected first level hyponyms of German Base Concepts for which we could not establish a synonymic link (synonym or near-synonym relation) to the ILI. We concentrated on candidates

- which are quite familiar and frequent in German;
- which are important concepts for further substructuring of semantic domains,
- which represent coding gaps in the ILI (whose co-hyponyms are more or less completely listed within WordNet 1.5).

For all the potential new ILIs that we have proposed, we found equivalents or at least glosses in English or bilingual lexicons, eg. the HarperCollins MRD German--English.

Within our proposed set we have not focused on concepts

- which are pointing to cultural gaps, eg. due to different education and administration systems (which are domains of very different lexicalizations in Germany versus in GB or USA). The only concept proposed in this regard is *Lebensgefährte*'companion through life'.
- which have multiple (non--synonymic) equivalence links;
- which constitute complex compounds like *Lebensrettungsgesellschaft* ;

We proposed 155 candidates for potential new ILIs, 145 nouns and 10 verbs. 40% of the verbs are from the field of creation verbs, eg. *töpfern*'do pottery'.

2. Some examples

2.1 Familiar and frequent concepts:

Fundort'place where something is/was found'
Fahradweg'cycleway' (should be quite common in Dutch too 'fietspad')
Zeitmangel'lack of time'
Kontonummer'account number'
Meisterschaft' championships'

2.2 Abstract expressions and (natural) kind terms which dominate a lot of hyponyms like

Hartkäse'hard cheese'
Leichtmetall'light metal'
Kernobst 'malaceous fruit'
Zierpflanze'ornamental plant'

2.3 Missing measure units (similar units have been provided):

Quadratkilometer'square kilometre'
Pfund'pound' (500 gram)

2.4 Missing expressions whose cohyponyms are more or less fully listed:

There are many different kinds of 'pressure', but for instance 'tyre pressure' (*Reifendruck*) and 'excess pressure' (*Überdruck*) are not encoded.

2.5 Some new countries in eastern Europe and their respective nationalities are not yet encoded like ‘Czech’, ‘Slovakia’, ‘Bosnian’ etc.

3. Proposals and Perspectives

Because of the diverging selection strategies and criteria being applied by the different sites (eg. with automatic methods or manually), it could be very useful to exchange the potential new ILIs that have been proposed by each site among the partners and to decide whether the concept in question is eventually quite essential or common in another language, too. This could increase the amount of the overlap across the EuroWordNet languages.

The tendency of over-representing English and American pecularities of lexicalizations and cultural conditions in EuroWordNet can thus be minimized.

For political correctness, new Eastern European countries and nationalities, respectively, should be fully accounted for as well as new European currencies (Euro)

3.4. Selection of the Italian candidates for new ILI-records.

The work aimed to circumscribe a subset of Italian candidates to be new ILI records has been carried out starting from the extraction of all the Italian synsets that have not been mapped to the ILI by means of a synonymy or near_synonymy equivalent relation. During the semi-automatic mapping phase, the synset is first compared with the lemmas in the bilingual source and then mapped via some equivalent relations, according to its position respect to the Italian and English taxonomies.

A synset could not be linked to an appropriate equivalent for different reasons:

1. the Italian lemma is not present in the bilingual source;
2. the Italian lemma is present but not the sense that we are trying to map;
3. the lemma and the sense are present in the bilingual source but we couldn’t find an equivalent in WordNet1.5.

For what is concerning to the case n. 1 and 2, the reasons for the absence of information could derive from either the fact that the word simply hasn’t been inserted in the bilingual source (as lemma or sense) or because the translation into English doesn’t exist.

The most interesting case seems to be the third one and, in order to circumscribe these synsets, we proceeded extracting the entries that, even though mapped via en eq_has_hyperorym relation, were present as lemmas in the bilingual source. Unfortunately, the obtained information is not disambiguated and contains also the case we have before explained in point two.

Maybe it would have being possible obtaining a partly disambiguated result, writing an extraction procedure that worked in the opposite way compared to the mapping procedure and that performed a comparison not only at the lemmas level but also at the level of the hierarchical assignment. It has been judged too complex and onerous under the point of view of resources as well as time, therefore we have preferred taking into account a phase of manual revision of the output.

Following this method, we have obtained a first list of Italian nouns and verbs, followed by the translation into English found in the bilingual source and by the file_offset of the hyperonym:

```
|pappa 1|n|11263|
TRADUZIONE: baby_food  mush  porridge
|pappa| |baby_food| |11263|
|pappa| |mush| |11263|
|pappa| |porridge| |11263|

|mandorlato 1|n|4879808|
TRADUZIONE: nut_brittle
|mandorlato| |nut_brittle| |4879808|

|lodare 3|v|530290|
TRADUZIONE: praise   to_praise
|lodare| |praise| |530290|
|lodare| |to_praise| |530290

|accecare 1|v|772512|
TRADUZIONE: blind  dazzle  to_blind  to_dazzle
|accecare| |blind| |772512|
|accecare| |dazzle| |772512|
|accecare| |to_blind| |772512|
|accecare| |to_dazzle| |772512|
```

With this information it has been possible to create the first part of the required record in the "new ILI" format. In order to assign the information regarding the link to the Top Concepts of the EWN Ontology, we have performed a comparison between the "file_offset" field of the obtained entries and the same field in the Ontology file.

We have assigned the correspondent Top Concept in case there was direct "hyperonymy relation" between the synset and the Top concept, otherwise the link have been established following recursively and vertically the hyperonymical chain. In this way we have obtained a file with about 1000 potential new ILI synsets. A long phase of manual revision of the results has been necessary since the information was not disambiguated. We also had to verify the sense number of the suggested new English entry.

Some taxonomies, above the others, highlighted the presence of lexical gaps in WordNet1.5, generally taxonomies pertaining to some very specific concrete nouns (in the Italian wordnet very specific senses are present, since the extraordinarily flat original hierarchies), like, for example, many kinds of fabrics (sangallo –broderie_anglaise-, loden –loden-), shops and factories (camiceria 1 and 2, shirt-shop and shirt-factory), kinds of clothes (spezzato –coordinated_jacket_and_trousers-) ecc..

It's clear that many of these records pertain to word meanings that the bilingual resource translates only by means of multiwords of two or more words, and it shows that the Italian sense has not been properly lexicalized in English. For example:

olimpionico (competitor_in the_Olympics), motopeschereccio (motor_fishing_vessel), capannello (knot_of_people), entrobordo (boat_with_an_inboard_engine), salumificio (cured_pork_and_meat_factory) ecc..

Only a little part of this record has been manually revised (about 60 nouns and 70 verbs records) and it's difficult to infer some real conclusions having analyzed so few data. We think that there's no need to increase the number of ILI records if we can establish some composite, complex links to the ILI. For example, in WordNet we couldn't find an equivalent for the Italian word "guerriglia" (guerrilla, guerrilla warfare) but we codified the entry with an eq_involved relation to the synset {guerrilla, guerilla, irregular, insurgent} (a member of an irregular armed force that fights a stronger force by sabotage and harassment) and with an eq_has_hyperonym relation to the synset {battle, fight, engagement}:

```
0 @18560@ WORD_MEANING
 1 PART_OF_SPEECH "n"
 1 VARIANTS
  2 LITERAL "guerriglia"
   3 SENSE 1
   3 EXTERNAL_INFO
    4 SOURCE_ID 2
 1 INTERNAL_LINKS
  2 RELATION "has_hyperonym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "battaglia"
     5 SENSE 1
  2 RELATION "involved_agent"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "guerrigliero"
     5 SENSE 1
 1 EQ_LINKS
  2 EQ_RELATION "eq_has_hyperonym"
   3 TARGET_ILI
    4 PART_OF_SPEECH "n"
    4 WORDNET_OFFSET 527805 ← battle, fight, angagement
  2 EQ_RELATION "eq_involved"
   3 TARGET_ILI
    4 PART_OF_SPEECH "n"
    4 WORDNET_OFFSET 6122095 ←guerrilla, guerilla, irregular, insurgent
```

In the same way, we built the complex link for the Italian synset “insabbiamento” (silting up) using an eq_is_caused_by relation to the WordNet verb {silt_up}:

```
0 @18844@ WORD_MEANING
 1 PART_OF_SPEECH "n"
 1 VARIANTS
  2 LITERAL "insabbiamento"
   3 SENSE 2
 1 INTERNAL_LINKS
  2 RELATION "xpos_near_synonym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "v"
    4 LITERAL "insabbiare"
     5 SENSE 1
  2 RELATION "has_hyperonym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "conseguenza"
     5 SENSE 1
  2 RELATION "has_hyperonym"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "atto"
     5 SENSE 1
  2 RELATION "involved_instrument"
   3 TARGET_CONCEPT
    4 PART_OF_SPEECH "n"
    4 LITERAL "sabbia"
     5 SENSE 1
 1 EQ_LINKS
  2 EQ_RELATION "eq_is_caused_by"
```

3 TARGET_ILI
4 PART_OF_SPEECH "v"
4 WORDNET_OFFSET 849843 ← silt up

So far, this work has not been done systematically but we think that this method could be used in the future to codify many Italian synsets for which we couldn't establish an appropriate eq_(near_)synonym relation.

3.5. Selection of the Estonian candidates for new ILI-records.

The concepts included in the Estonian wordnet for which there did not exist adequate equivalences in ILI records can be loosely divided in three groups.

First, there is a number cases where Estonian because of its dominating agglutinative structure and of the large use of compound words (that orthographically are one word - just as e.g. in German), has lexicalized concepts (meanings) which in English are expressed by collocations or (syntactic) compounds. Since these concepts form a normal part of the lexical system of Estonian, we had to include them in our wordnet.

Secondly, according to the Project plan we did an "in-depth" analysis of some domains which are not represented in ILI (=WordNet 1.5) at the same depth (measuring instruments, musical instruments, communication). This work brought in a large number of quite specific concepts.

Thirdly, there is a number of concepts (meanings) in Estonian that can be explained by our specific cultural background that differs from that of the typical Anglo-Saxon or even Western one in general that is represented in the WordNet.

Let us characterise these types in more detail.

1. Language-specific concepts.

1.1. In Estonian derivational means (suffixes) are often used to produce new words with new meanings which have fixed relationships to the meanings of words that constitute the derivational basis. For instance:

Lükand (from lükkama – to push) – "anything that works as a sliding door or window"; *nd* is a productive suffix;

Hapendama (from hapu – sour) - "to make/turn sour, to let get sour (e.g. milk)

Määrduma - "to become dirty" (cf. *määrima* – to soil, to smear; *du*- is a reflexive suffix).

Hilinema – "to be late " (cf. hilja - late).

1.2. Estonian makes large use of orthographical compound words and their meanings again are included – often as hyponyms of more general terms – into the lexical system . For instance:

Draamanäitleja - actor or actress who plays in dramatic genre (hyponym of *näitleja* – actor or actress)

Filminäitleja - actor or actress who plays in movies

Kaubandustöötaja – person whose occupation is commerce or trade (*töötaja* – worker, *kaubandus* – commerce)

Kättemaksuplaan – plan of revenge (hyponym of *plaan* – plan)

2. That the in-depth analysis of certain domains brought in new concepts not present in ILI should be apparent and does not need any specific comments. Some examples from the domain of measuring instruments:

Dendromeeter - dendrometer, a measuring instrument used for measuring the diameter or perimeter of trees

Butyrometer – butyrometer, an instrument used to measure fat in the milk

But also:

Lauakaal – "table scale", a (small) instrument used for weighting (food etc.) on the tables

Postikaal – weighting device used in post services

3. As examples of this third type, where the most natural explanation seems to be the difference in the cultural background, we may present first such examples as:

Aastaplaan – "one-year-plan", a plan of action for one year used, first of all, in planning economic activities (aasta – year)

Eksamiprotokoll - report of examination results (in the universities)

But e.g. also

Lõõtspill – a portable musical instrument with bellows and a keyboard (*lõõts* – bellows)

Seljanka - thick Russian meat or fish soup

Hingedeaeg – (hing – here "ghost", aeg – time, period), the darkest period before the winter solstice when ghosts were believed to visit homes

Maavanem .- head of *maakond*, the largest administrative unit in Estonia.

Of course, not all concepts for which there does not exist an eq-synonym in ILI are exhausted by this classification. For instance, as in every language, also in Estonian there is a lot of words that e.g. descriptively represent some object, situation or activity. This descriptive part constitutes a part of the meaning of the corresponding word and often it is difficult to give an equivalent in another language. Or there are situation where different languages lexicalise different individual concepts. These kinds of discrepancies have been discussed in other deliverables.

At last, it must be noted that there are cases where there seems to be a normal eq-synonymy relationship between an Estonian concept and an ILI concept, especially when we take into account e.g. hyponyms and hyperonyms also, but the ILI gloss makes one to hesitate if we really have to do with the same concept. For instance, the Estonian concept *grammatiline kategooria* seems to correspond neatly to the ILI category *grammatical . category* when we look at its hyponyms (*case*, *gender* etc). But the gloss says "a category of words having the same grammatical properties". In Estonian grammatiline kategooria does not mean a "a category (=a class) of words", it means the corresponding grammatical properties that characterise a (class of)words (such as case or gender).

4. Comparison of the proposed ILI-concepts

The proposed new ILI-records are compared in two different ways. First, we did a global comparison of the Dutch, German ands Italian sets to see how much overlap there is and to what extent it may be possible to do an automatic mapping. Since the potential overlap is very low, we have also checked manually whether the proposed concepts (from the German and Italian wordnet) do occur in one specific other wordnet or language, in this case the Dutch wordnet. This gives an indication of the potential overlap if the proposed sets would have been more comprehensive. Both comparisons are discussed in the next two subsections. In the third subsection, we describe the opposite effect that the lexicalization in non-English languages suggests a reduction of concepts in the ILI rather than an extension.

4.1. Global comparison of the proposed ILIs

We carried out a comparison of 3 of the above selections. These selections have been compared in terms of:

1. the proposed English variants
2. the variants in the local Languages
3. the glosses
4. the references to the ILI by the complex equivalence relations

Table 1: Proposed collections of ILI candidates

	Nouns			Verbs		
	NL	DE	IT	NL	DE	IT
ILI Candidates	593	145	78	93	10	67
Variant Tokens	593	145	78	93	10	67
Variant Types	551	144	77	82	10	66
Source Variant Tokens	756	228	78	121	14	78
Source Variant Types	731	224	75	93	14	73
ILIRef Tokens	678	146	71	122	10	81
ILIRefTypes	491	109	45	99	9	22
Gloss	593	14	56	91	0	67

Table 2: Intersections of ILI references and ILI variants

	ILI REFs (mostly hyperonyms)		ILIVars	
	Nouns	Verbs	Nouns	Verbs
NL	491	99	551	82
DE	109	9	144	10
IT	45	22	77	66
NL&DE	10	0	2	0
NL&IT	6	3	1	0
DE&IT	5	1	0	0
NL&DE&IT	3	0	0	0
Union Intersections	15	4	3	0

The intersection of the English variants is very low (3 for nouns and 0 for verbs), even the intersection of the ILI-references is low. These ILIs are mostly referred to by an EQ_HAS_HYPERONYM relation. This means that there is not much common ground for a comparison. The general conclusion is that the sets are too small to give a significant result.

Intersecting Noun Variants

```
((bar))
((keyboard instrument))
((spinning mill))
```

In quit a few cases, like “bar”, the entry of the proposed variant does occur in WordNet. In the case of the Dutch wordnet this is because the variants are automatically generated from the translation fields. Often a more general

word is given when there is no good equivalent in English. The more general word then still matches with a wordnet entry. The next table gives an overview of the proposed variants that match wordnet entries:

Table 3: Overlap of entries with WordNet1.5.

	Nouns			*Verbs*		
		Not In WN	In WN		Not In WN	In WN
NL	551	288	263	82	32	50
DE	144	132	12	10	9	1
IT	77	66	11	66	63	3
Union		484	285		104	54

There may be higher matching because many variants are multi-words. In that case, the use of hyphens and spaces may differ. In all these cases, we have to inspect manually whether the proposed sense is included or not. It would be better if these concepts are explicitly marked as new senses.

4.2. Evaluation of the potential overlap

To get an idea what the possible candidates can be, we have inspected a sample of the Italian and German ILI-records to see if they potentially could overlap with Dutch synsets. A closer look at the concepts showed that a larger set of proposed ILIs would certainly increase the overlap.

Comparison with 36 proposed German ILI records showed that 50% of the nouns (18) have an equivalent in Dutch, but these were either not in the selected subset or not in the Dutch source.

Examples:

Arbeitszeitverkürzung	arbeidstijdverkorting	(shortage of working hours)
Campingzubehör	kampeeruitrusting	(camping equipment)
Darmkrebs	darmkanker	(cancer of the bowel/intestine)

For 59 proposed Itialian ILI-nouns there is at least an overlap of 30% (20) with Dutch.

Examples:

baita	berghut	(cabin in the mountain)
bobbista	bobber	(person who rides a bobsled)
contraindicazione	contra-indicatie	(counter-indication)

For a few nouns we believe there is an equivalent in WN1.5, for example 'Barren' (bar, measure unit as of gold), namely 'ingot' or 'bullion'. In many other cases, we could not judge as non-natives what concept was intended. Expert knowledge of the languages is needed to define the precise overlap. In most cases we based our judgement on the glosses, but this was not always possible. For instance for the following examples it is not clear to us what is meant:

	Gloss
'maschera',	stock character
'ghiera'	ring nut

This means that not only expert knowledge is needed but also that the proposed ILI-concepts need to be specified more (e.g. by providing more complex equivalence relations).

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. Presumably, two or three verbs are equivalent in Dutch. Again the problem here lies in the gloss, it is not clear whether 'fräsen', gloss: *mill-cut* is the same as the Dutch 'frezen' which was assigned to an equivalent in WN1.5 ('mill').

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 are multi words in Dutch.

abbuiarsi	get serious
imbarbaririsi	become less civilized
incattivire	make wicked
infiacchire	make lazy

Many of these cases 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. This would reduce at least half of the proposed ILI-verbs. 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.

It is clear from this comparison that at this stage it is not possible to know for sure what the complete overlap is. In many cases it is difficult to decide even manually, let alone, to do this automatically. An automatic matching may support a manual process. Nevertheless, much more detailed information has to be provided, possible using other bilingual dictionaries, such as Dutch-German, Dutch-Italian. Only for a subset of ILI's we can decide that they should be added.

There are a few classes of concepts that we would like to exclude from being added to the ILI:

- names for instances;
- Xpos synonyms: in some cases there is an equivalent with a different POS;
- (semi-)productive patterns of lexical incorporation, e.g. “to become mad”;

Although instances can be added to the ILI, we think it is not the highest priority to add these. Many facts and instances can be derived from various sources. There is no reason why we should add some and not others. An exception to this may be nouns that denote the names of a country or the inhabitants. The second class of xpos synonyms does not represent new ILIs. The concept is present but referred to by a word with different POS. A normal equivalence link can be created to the ILI. The third group is most difficult to delineate. It includes compound nouns and verbs that can often be derived systematically. In these cases, the concept does occur in other languages/cultures but is expressed differently.

Genuine additions to the ILI are represented by so-called cultural gaps: the English culture does not distinguish the concept and it is therefore not lexicalized. This is not always easy to decide, and we cannot reverse this hypothesis either: the fact that lexicalizations are similar does not mean that the concepts are shared. For example, there are a few cases of dishes or food among the proposed ILIs, where it is unclear that we are dealing with new dishes with the same name, or that the differences are not really relevant. For example, “Bauernbrot” could be translated as ‘brown bread’. There can be differences in ingredients but it is unclear how relevant these are.

It is thus still a major problem how to find criteria to differentiate between the above classes of gaps, and how to develop (preferably automatic) techniques to select them. Currently, we can only extract incidental examples of new ILIs from the proposed set. More fundamental research is needed to get more results.

4.3 Multiple links between language specific wordnets and the ILI

This section describes an investigation into up to which extent the sense granularity displayed by English lexicalizations of ILI concepts can be reduced by taking mappings into account between language wordnets and the ILI. In a number of cases one language specific concept has been linked to multiple ILI concepts by means of EQ_SYNONYM and EQ_NEAR_SYNONYM relations. These multiple mappings have been evaluated with respect to their validity as sense clusters if the WordNet1.5 synsets that lexicalize the concepts share the same word form. This investigation is part of a larger set of strategies applied to clustering WordNet1.5 sense distinctions. These have been described in deliverable D2004.

The data used originates from the Dutch and Italian wordnets, and cover nouns only. The Dutch data consists of 931 cases where there is a link between one local concept and multiple ILI records. For 73 of these two or more of the synsets involved share the same word form in WN1.5. The Italian data contains 342 possible groupings, of

which 30 contain synsets involved share the same word form in WN1.5. These potential WordNet1.5 clusters have been selected for evaluation, because they allow us to clusters senses of the same word.

A manual and therefore inescapably subjective evaluation was performed on them. This resulted in a set of 5 valid clusters on the basis of Italian equivalence links. 4 (26%) of these are already covered by the existing ILI clusters (see deliverable D2004). 56 valid clusters were selected on the basis of Dutch equivalence links of which 7 (25%) are already covered by clusters in the present ILI.

4.3.1 A semantic typology of multiple equivalence links.

An important observation that must be made here is that using the sense distinctions of existing resources will always cause a mapping problem between the various resources because different resources depend on different lexicographic practices, space limitations etc.

In cases where several ILIs offer plausible conceptual equivalence links with one language specific concept, there are several possible explanations:

1) The ILI concepts are not sufficiently enough distinguishable from each other. This allows us to cluster the ILI concepts on the basis of generalization:

welwillendheid 1
- a disposition to kindness and compassion; benign good will: "the victor''s grace in treating the vanquished" (*good will* 1)
- disposition to do good (*benevolence* 1)
- an inclination to do kind or charitable acts (*benevolence* 2)

2) The ILI concepts are distinguished enough, but the sense distinctions offer different perspectives on the same meaning (logical metonymy) or display instances of systematic polysemy:

zonsondergang
1 the time in the evening at which the sun begins to fall below the horizon (*sunset* 1)
2 atmospheric phenomena accompanying the daily disappearance of the sun (*sunset* 2)
3 the daily event of the sun sinking below the horizon (*sunset* 3)

government
1 the organization that is the governing authority of a political unit
2 the body of persons who administer something

3) the sense distinction in the local wordnet is not fine-grained enough, and the linked ILIs function as subsenses of the local sense.

This does not necessarily allow us to cluster the ILIs, because the resulting ILI cluster does not reveal the right level of semantic underspecification. Take the following example taken from the Dutch data:

l*ijden* 1
suffering
3 psychological suffering
4 physical suffering

One could contend that there is a genuine difference between the two sorts of pain, as many may acknowledge from experience, and therefore the clustering would be invalid in this case.

fout 2
error
1 a wrong action attributable to bad judgment or ignorance or inattention
6 a deficiency

A deficiency is not necessarily the result of ignorance/inattention/bad judgement.

4) there is a genuine difference between the conceptual structures of the languages involved.
This can be expressed by a higher degree of specialized lexicalization in one language:

Stad
- a large and densely populated urban area; may include several independent administrative districts; "Ancient Troy was a great city" (city 1)
- an incorporated administrative district established by state charter; "the city raised the tax rate" (city 2)
- an urban area with a fixed boundary that is smaller than a city; "they drive through town on their way to work" (town 1)

5) the WordNet glosses and ontological embedding are not good enough for proper evaluation.

state of mind
1 the state of a person's cognitive processes
2 a temporary psychological state

These different cases can be used for a typology of the status of ILI concepts. For instance, the ontological status of ILI records as language independent concepts can be differentiated along criteria such as logical metonymy and subsenses with/without different lexicalization patterns in the languages. Factoring out these semantic observations will contribute the definition of a more language independent interlingua as described in the next section.

4.3.2 Discussion
If we scale up to all multiple ILI references on the basis of these results, in which case we

- include the EQ_(NEAR_)SYNONYM clusters where the WN1.5 synsets do not share a word form, and
- presume that our manual evaluation of the potential cluster described above is the golden standard,

we can infer that the Dutch multiple links can be accepted as they are as sense clusters in the form of composite ILIs with 76% confidence, the Italian with 50% confidence.

To conclude, it needs to be mentioned that there is an interesting connection with the clusters distinguished previously (see deliverable D2004). Sense clusters derived on the basis of different strategies often show a considerable degree of overlap. They complement each other and, if taken together, form larger clusters encompassing a greater number of WordNet1.5 than the original clusters that constitute the so-called supercluster.

An example:

thought
1. the content of cognition; the main thing you are thinking about; "it was not a good idea"; "the thought never entered my mind"
2. the process of thinking (especially thinking carefully); "thinking always made him frown"; "she paused for thought"
3. the organized beliefs of a period or group or individual; "19th century thought" or "Darwinian thought"
4. a personal belief that is not founded on proof or certainty; "my opinion differs from yours"; "what are your thoughts on Haiti?"

Senses 1 and 4 had been clustered manually in a previous round. Senses 3 and 4 had been clustered automatically on the basis of the sorority principle. Further, the Dutch *gedachte* 1 is linked to 2, 3 and 4 in the evaluated clusters described above.

Consequently, all senses of thought have been clustered together in one generalization grouping on the basis of the accumulation of smaller clusters. This large cluster is underspecified for the notions *content* (as a mental object) versus *process* and *belief* versus *epistemological certainty*.

5. Towards a condensed and universal index of meaning

The unmatched concepts described in the previous section can be classified as follows:

1. Matches to different Part of Speech: synsets that can be matched to an ILI-record that has the same meaning but a different part-of-speech.
2. Exhaustive Links: synsets whose (productive) meaning is fully captured by several, possibly complex equivalence links to multiple ILI-records.
3. Incomplete Links: synsets that can only be linked to a hyperonym ILI record that classifies it.
4. Unresolved Links: cases that cannot even be linked to a hyperonym ILI record

The first category contains part of speech mismatches like the Dutch static verb “aanstaan” (be ajar) discussed above. 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 by means of an EQ_NEAR_SYNONYM relation. 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. 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). This is shown in Figure-1 below, where the box in the middle is the ILI with a Composite ILI-record “depart” that groups the verbal and nominal forms “depart” and “departure” respectively. To the left and right, we see the lexicalizations in Dutch and English which are similar. Below this example we see another case where the adjective “nice” and the verb “like” are grouped and the Dutch and English are different: i.e. there is only an equivalent for the adjective in Dutch.

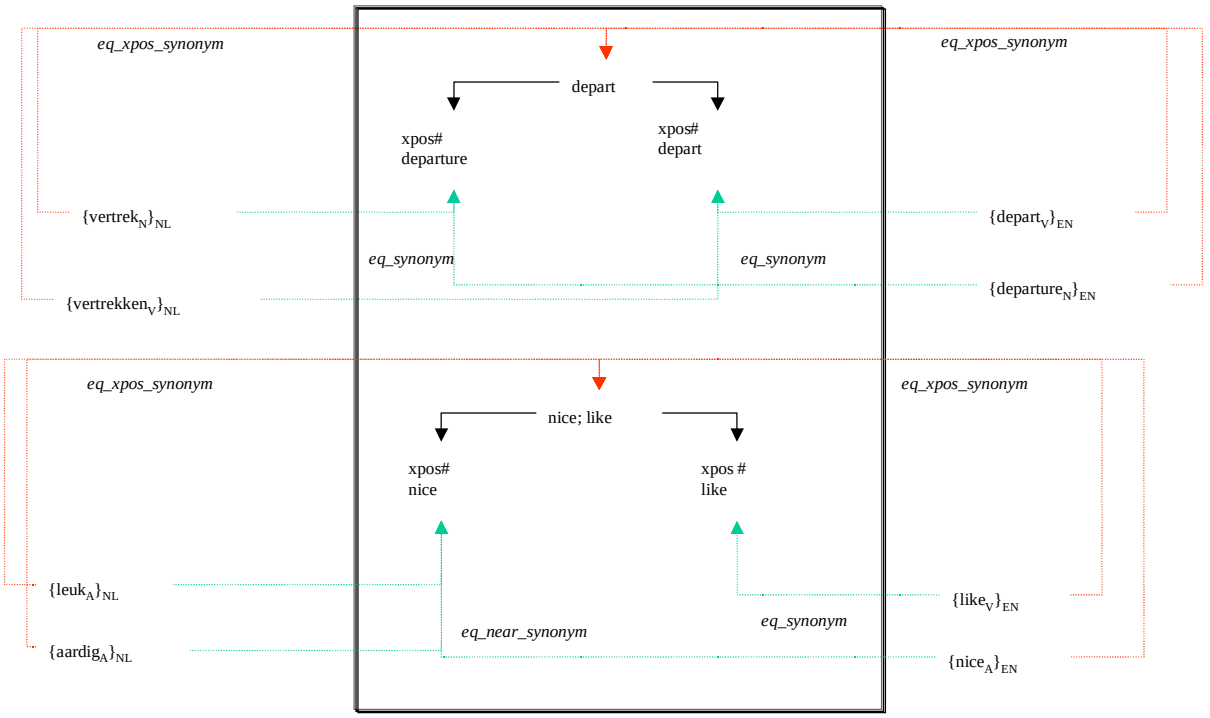

Figure-1: ILI with part-of-speech neutral Composite ILI-records.

The second category of synsets often follows a regular pattern of compounding or derivation. We have seen some examples of compound verbs which meaning is compositionally derivable from the compound structure, e.g. in Dutch:

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

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. 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 various languages. Finally, the concept is not unknown in English but is simply not lexicalized as a single verb and is not considered to be a lexicalized form.

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. In Figure-2 below, we see how these concepts can be linked to the ILI via multiple exhaustive links. Here we see first several cases of Dutch and German compound verbs that can be linked to the ILI exhaustively and uniquely, where the latter means that a single Dutch and German word linked in the same way can thus be assumed to be equivalent across the two languages, even when there is no connecting index record. Below the verbs are also some similar examples of nouns with lexicalizations in Dutch and Spanish.

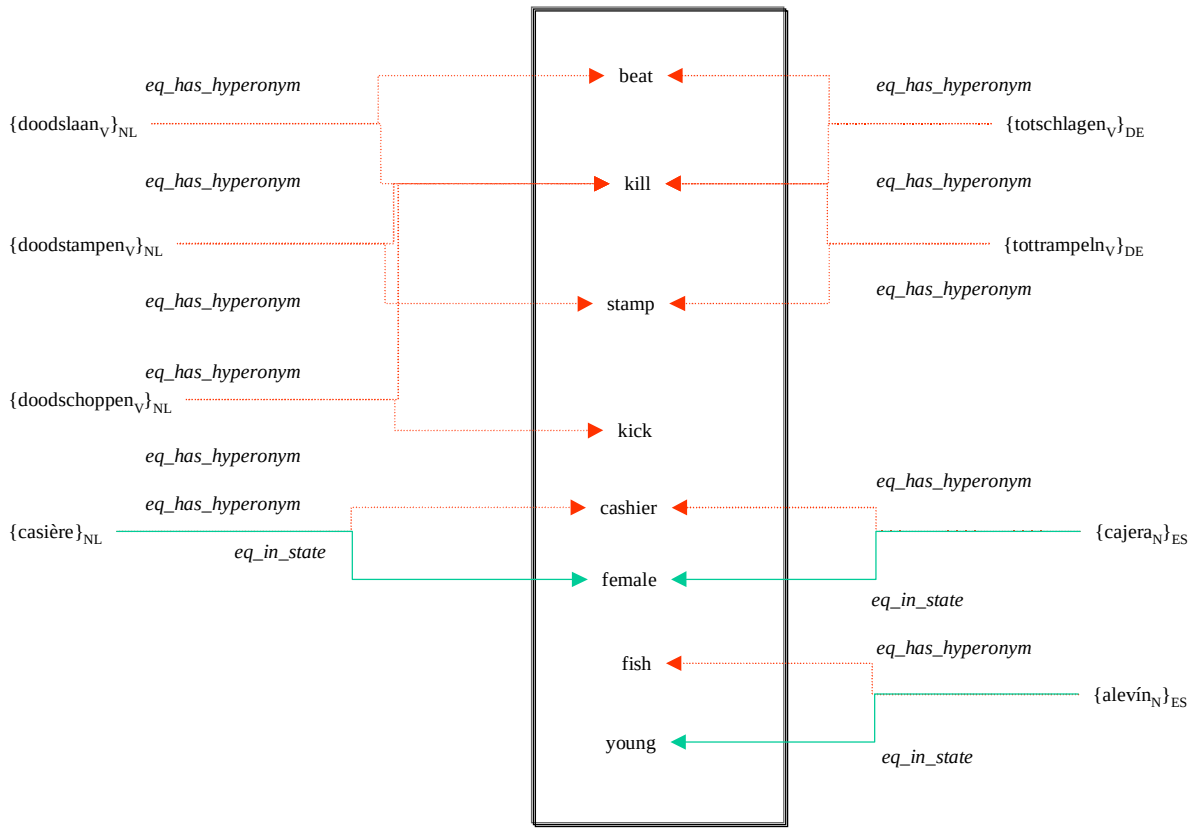

Figure-2: Productive meanings exhaustively and uniquely linked to the ILI

Just as with the cross-part-of-speech matches, the above strategy would imply that current ILI-records derived from WordNet1.5 that can be linked and predicted in the same way should be removed from the standardized list. Note that these word meanings will still be present in WordNet1.5 or WordNet1.6 for English, which is still linked to the ILI. Obviously, these English word meanings should be linked exhaustively via complex equivalence relations, as is the case for the other languages.

Category 3 and 4 may then contain genuine gaps, that either do not occur in the English/American culture or they are lexicalized but have been omitted from WordNet1.5. These can therefore be seen as source for genuine

new ILI-records. The estimates so far are that these form a minority of the unmatched cases. For most unmatched synsets, it is thus not really necessary to extend the ILI. Even stronger, 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 hyperonym 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 can be helped by looking at the morpho-syntactic markedness (e.g. regular compound structures), regular lexicalization patterns and corpus frequency.

Together with the grouping of senses in more global ILI-clusters (see, 2D004 Peters 1998), we thus see that we can reduce the ILI dramatically and incidentally extend it but still have a minimalized and condensed index of meaning that is universal. This is shown in Figure-3 below and further discussed in Vossen, Peters and Gonzalo (1999). However, it is also clear that this restructuring involves a lot of manual work covering many languages (from different language types) and detailed specifications of the meaning and exhaustiveness of the linking. Possibly, the differentiation of the concepts can also be achieved in the opposite way. ILI-records that are covered by many different wordnets can be seen as universal and thus belonging to the core of the index. Finally, the database will have to be extended to be able to express the differences in status of the meaning. Clearly, this cannot be done in the framework of this project anymore.

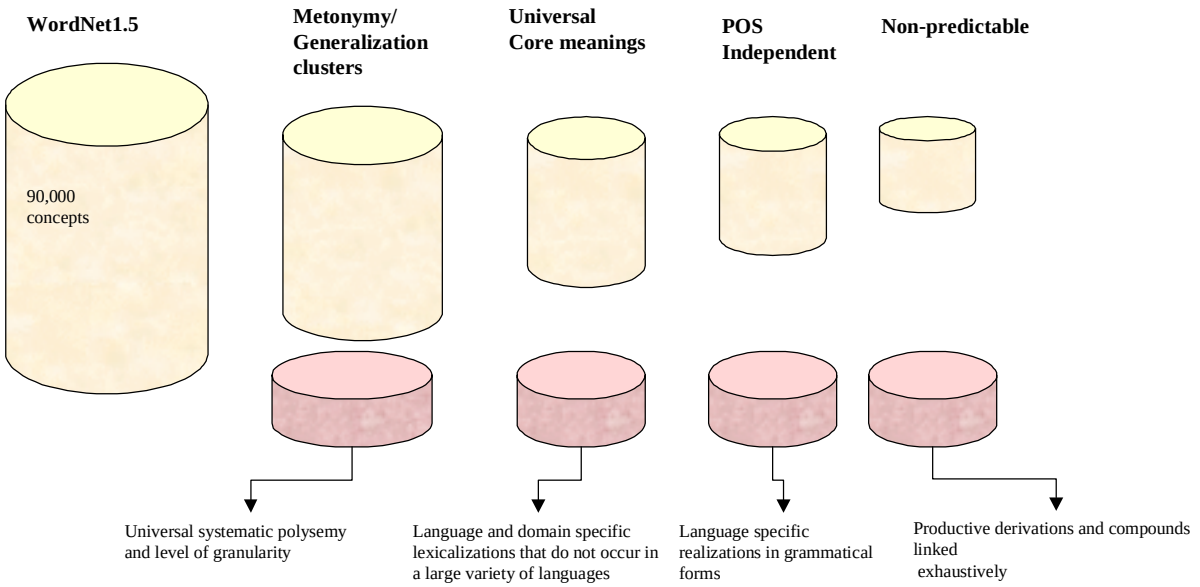

Figure 3: Towards a condensed, minimalized and universal index of meaning.

References

Fellbaum, C. (ed.)
1998 *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
Martin W. and J. Tops
1989 Groot woordenboek Nederlands-Engels. Van Dale Lexicografie. Utrecht.
Peters W.
1998 "Restructured ILI" EuroWordNet (LE 4003), Deliverable 2D004, University of Sheffield.
Vossen, P., L. Bloksma, P. Boersma
1999 The Dutch Wordnet, EuroWordNet (LE2-4003) University of Amsterdam.
Vossen, P., W. Peters and J. Gonzalo
1999 Towards a Universal Index of Meaning. In the proceedings of the ACL-99 Siglex workshop, University of Maryland.

Appendix I Example list of Czech nouns with missing equivalents

Explanation:
1. do not have WN 1.5 equivalent at all - this is marked by //
2. there is a WN 1.5 equivalent but with different sense(s) - marked as !
3. ?? - problematic cases, eg. probably not noun or have to be checked.
4. the numbers refer to the senses in Czech resource(s)

administrátor=2: (%farnosti ap.) !administrator in a catholic church,
adresa=4: (%pozdravná) address !letter of congratulation
adresáø=2: (%souborù na disku) //computer directory
agent=3: (%policejní) //secret police man, in WN - agent:4 federal agent ??
agentura=3: (%výzvìdná) intelligence service (agency in WN)
akce=1: (%èinnost) action, project, eq. is activity
alkohol=3: (%alkoholy) ?? (alcohol:3 in WN)
analýza=4: (%syntaktická) //parsing, //a parse
anonym=1: (%autor) //anonymous {+author,writer}
anonym=2: (%dopis) //anonymous letter, (%zlomyslný) //poison-pen letter
aparát=3: (%kritický) !apparatus
áèko=1: (%písmeno) letter A, (%známka) //first class (.mark), //A grade
áèko=2: (%struna A) //string A, //A minor
áèko=3: (%oznaèení) //label A
bahno=5: (%morální) !ordure, !muck
bahno=6: (%pomluvy) !mire, !muck
balast=2: (%slovní) //padding
balast=3: (%nezu•itkovatelné slo•ky) !ballast
balast=4: (%pøebyteèní pracovníci) dead wood !ballast
balet=3: (%soubor) !ballet, the ensemble
banka=2: (%krevní) !blood bank
banka=4: (%genová) !genetic bank (containing genes)
baroko=1: (%styl) !baroque style of architecture
baroko=2: (%stavba) !baroque building
barva=7: (%krev zvìøe) !colour, animal's blood
basa=2: (%pøepravka) !crate, !case, //box (of beers)
basista=2: (%hráè) bass player, //bassist
baant=2: (%pokrm) !pheasant (poultry meal)
baant=4: !urinal (a vessel)
beran=5: (%tvrdohlavý èlovìk) !ram {+pigheaded, wrongheaded, bullheaded} person
beseda=4: (%osvìtová organizace) //popular education organization
bezvìtøí=1: !calm (windless weather)
bezzemek=1: //landless (a person who does not own any land)
béèko=1: (%druhý, druhoøadý) //B, second-rate thing or person
béèko=2: (%oznaèení) //label B
béèko=3: (%nota) //B-flat
?? bìda=1: alas (not a noun)
bìh=4: (%postup, proces) !process, procedure, (%tok) !flow, may be eq_near_synonym
bìhoun=1: (%koberec) !runner (a kind of carpet)
bìec=3: (%souèástka, jezdec) !traveller, (*elekt.) !cursor
bilingvismus=1: !bilingualism
bláto=2: (%morální) !ordure, !muck
?? bod=2: (%mrazu ap.) freezing point - it is in WN!
bojištì=2: (%nepoøádek) //battle ground, here denoting a mess or disorder
bomba=3: (%ozaøovací) !bomb (radiation bomb in medicine)
bor=1: !pinewood
borovièka=1: (%strom) !dwarf pine tree
brambor=3: (%pole) //potato field
brána=4: (%mìstská) //town gate, //city gate
brigáda=2: (%prázdninová) //holiday job, //summer job, (%vedlejší pracovní pomìr) //temporary job
briliant=1: //brilliant, precious stone
brouzdaliště=1: paddling pool
brusiè=1: !grinder; !cutter
brynda=1: (%nápoj) //rot-gut, !swipes (a drink)
brynda=2: (%káva) //wet and warm (coffee)
bøeèka=2: (%špatný nápoj) //rot-gut, !swipes
budiknièemu=2: (%moula) !zombie (unable person)
bylina=2: (%léèivá) //medicinal herb
byrokrat=2: (%zkostnatìlý) !bureaucrat (a person who is not flexible)
pou=4: (*BrE) (%ka•doroèní slavnost v Anglii na památku sv. patrona,
kterému je zasvìcen kostel) !wake

Appendix II Computer terminology in EuroWordNet, grouped by part-of-speech (a, n, v)

a alphanumeric
a compatible
a deterministic
a graphic
a hierarchical
a incompatible
a interactive
a lossless
a online
a quick-and-dirty
a remote
a text-based
a under construction
a virtual
n A:
n abduction
n abort
n accelerator
n accumulator
n adapter
n address
n administrator
n agent
n alias
n alt
n anchor
n animation
n annotation
n Anonymous FTP
n anti-aliasing
n applet
n application
n architecture
n archive
n argument
n array
n assignment
n audio
n authentication
n automaton
n back door
n backbone
n background
n backtracking
n backup
n bandwidth
n banner
n basename
n Basic
n batch
n Baud
n BBS
n benchmark
n BIOS
n bit

n bitmap
n block
n bookmark
n bool
n bottleneck
n browser
n buffer
n bug
n bullet
n bus
n byte
n c++
n cache
n callback
n capacity
n central processing unit
n character
n character set
n checkbox
n checksum
n clickable image map
n client
n client-server
n clipboard
n cluster
n code
n command
n communication system
n compiler
n compression
n computer
n controller
n core
n coupling
n cryptography
n current
n cyberspace
n cylinder
n daemon
n data
n data processing
n data traffic
n database
n database front end
n debugging
n decryption
n dedicated line
n default
n delimiter
n design
n device
n dial-up account
n DIN
n directory
n disk
n disk drive
n display
n DNS
n domain
n domain name

n download
n driver
n DSP
n EBCDIC
n electronic mall
n electronic storefront
n ELF
n Emacs
n email
n embedded hyperlink
n emoticon
n emulation
n encapsulation
n encryption
n engine
n environment
n escape
n facsimile
n fallback
n FAQ
n FAT
n feedback form
n filename
n filtering
n finger
n firewall
n firmware
n flame
n flat file
n floating-point
n floppy disk
n focus
n font
n foreground
n Form
n form support
n format
n forwarding
n fragment
n freenet
n freeware
n FTP
n function
n gain
n gateway
n GIF
n gigabyte
n GNU
n Gopher
n graphics card
n gray-scale
n GUI
n guru
n hack
n hacker
n handler
n hard disk
n hardcopy
n hardware
n hexadecimal

n history
n home page
n host
n hotlist
n hotspot
n HTML
n httpd
n hyperlink
n hypermedia
n hypertext
n I/O
n icon
n IFF
n imaging
n impact printer
n index
n indexing
n induction
n inference
n information packet
n inline image
n instantiation
n integrated circuit
n integration
n integrity
n interface
n internationalisation
n Internet
n Internet account
n Internet service provider
n interpreter
n IP
n IP address
n ISDN
n Java
n JIT
n job
n jpeg
n jumper
n key
n keyboard
n keyword
n kilobyte
n laser
n laser printer
n leased line
n LED
n line printer
n link
n Linux
n LISP
n local area network
n log
n log file
n logging
n logic programming
n login
n logout
n Macintosh
n magnetic disk

n magnetic tape
n mail-bomb
n mail-filter
n mailbot
n mailbox
n mailing list
n mainframe
n majordomo
n mapping
n markup
n media
n megabyte
n memory
n menu
n microprocessor
n MIME
n minicomputer
n MIPS
n mirror
n mnemonic
n mode
n modem
n moderated mailing-list
n module
n monitor
n mouse
n MS-DOS
n multicast
n multiplexing
n multithreading
n net
n netiquette
n netizen
n newline
n newsfeed
n newsgroup
n newsreader
n noise
n number crunching
n octet
n OEM
n openwindows
n operating system
n operator
n option
n OS/2
n output
n overhead
n overloading
n overriding
n packet
n paging
n parallel processing
n parity
n partition
n Pascal
n password
n pathname
n PC
n PDF

n peripheral
n Perl
n pipeline
n pixel
n platform
n plotter
n point
n pointer
n POP
n port
n POSIX
n postmaster
n preference setting
n premastering
n primitive
n printer
n process
n programming language
n protocol
n provider
n proxy
n punched card
n queue
n radio button
n RAM
n real-time chat
n recursion
n redirection
n registry
n relational database
n release
n remote login
n requirement
n resolution
n resource
n ring
n ROM
n router
n scan
n script
n SCSI
n search engine
n segment
n semantics
n semiconductor
n sendmail
n server
n server-side include
n service
n session
n shareware
n shell
n shell account
n signature file
n SLIP/PPP
n SMTP
n software
n sound player
n source code
n spamming

n specification
n spooler
n spreadsheet
n SQL
n stack
n standard
n storage
n stream
n streaming
n subclass
n subdirectory
n subnet mask
n subroutine
n switch
n syntax
n system call
n tag
n tagging
n tape
n TCP/IP
n telnet
n terabyte
n terminal
n testing
n text
n text editor
n text-based browser
n throughput
n timeout
n token
n tool
n traffic
n transistor
n troff
n Trojan horse
n turn-key
n UNIX
n upgrade
n URL
n Usenet
n user
n username
n VM
n web
n webmaster
n what's new
n whitespace
n wildcard
n word processor
n workaround
n workstation
n WYSIWYG
v abort
v boot
v bootstrap
v click
v crash
v cross-post
v default
v delete

v dereference
v download
v drag
v emulate
v encrypt
v format
v hit
v input
v instrument
v interrupt
v load
v log off
v log on
v navigate
v post
v read
v reboot
v spool
v surf
v swap
v unzip
v upgrade
v upload
v zip