EuroWordNet as a Multilingual Database

PIEK VOSSEN

Universiteit van Amsterdam,

Faculteit Geesteswetenschappen,

Spuistraat 134,

1012 VB Amsterdam,

The Netherlands

E-mail: Piek.Vossen@hum.uva.nl

Http: //www.hum.uva.nl/~ewn

Abstract

EuroWordNet builds a multilingual database with wordnets for several European languages. Each language specific wordnet is structured along the same lines as WordNet (Miller et al. 1990): i.e. synonyms are grouped in synsets, which in their turn are related by means of basic semantic relations, such as hyponymy, meronymy, semantic roles. Each wordnet uniquely describes the lexicalization pattern of a language. Multilinguality is achieved by storing the language-specific wordnets in a central lexical database in which equivalent word meanings across the languages are linked to a so-called Inter-Lingual-Index (ILI). In this paper, we will address the way multilingual relations are expressed in EuroWordNet and how they are (semi-)automatically extracted from the monolingual wordnets and bilingual dictionaries. Different kinds of equivalence mappings are described to deal with fuzzy mappings and gaps. The fact that these equivalence relations are established at a more global synset level and that it is possible to browse to closely related synsets, makes it possible to get a comprehensive conceptual match of concepts across languages, even when lexicalizations differ. Other evidence, such as co-occurrence probabilities and morpho-syntactic constraints, can then be used to find a translation or a correct combination of words in a target language that is appropriate in a context.

Key-words: Multilingal Database, Semantic Networks, Automatic Equivalence Matching

1. Introduction

EuroWordNet is an EC-funded project (LE2-4003 and LE4-8328) that builds a multilingual database with wordnets for several European languages. Currently, the project covers the following languages: English, Dutch, Italian, Spanish, French, German, Czech and Estonian.[1] Each wordnet is structured along the same lines as WordNet (Miller et al. 1990): i.e. synonyms are grouped in synsets, which in their turn are related by means of basic semantic relations, such as hyponymy, meronymy, semantic roles. Each of the wordnets is a language-specific structure, uniquely expressing the lexicalization relations between the words of a language (Vossen 1998: 73-89). The wordnets contain general vocabulary and the conceptual coverage is balanced over a common set of Base Concepts. These Base Concepts are selected for their importance in a variety of wordnets (Rodriguez et al. 1998: 91-115). The size of each wordnet will be between 15-30K synsets, representing 30-50K word-senses, and 10-20K lemmas.

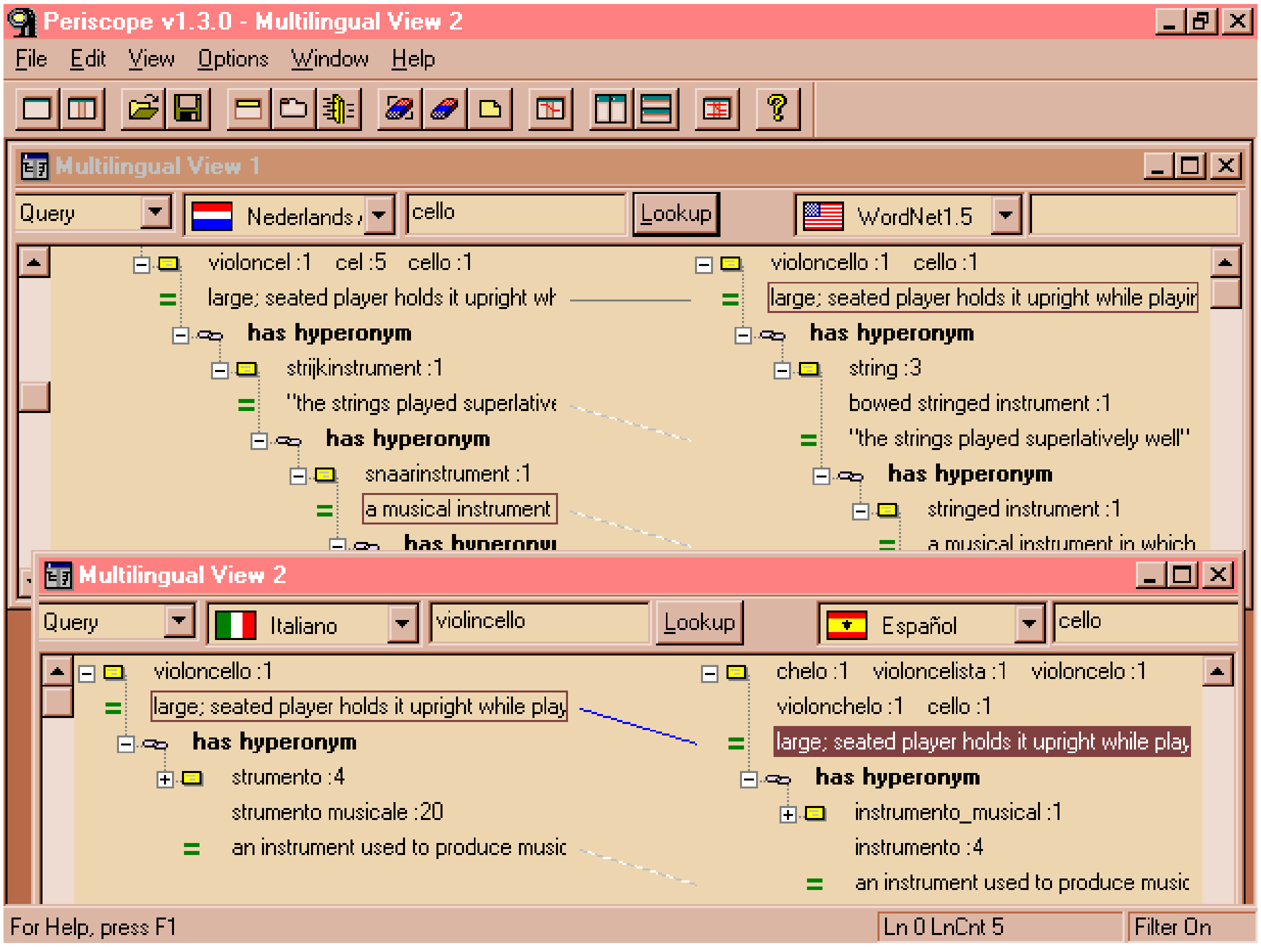
Multilinguality is achieved by storing the language-specific wordnets in a central lexical database in which equivalent word meanings across the languages are linked via a so-called Inter-Lingual-Index (ILI). In this paper, we will address the way multilingual relations are expressed in EuroWordNet and how they are (semi-)automatically extracted from the monolingual wordnets and bilingual dictionaries. In the next section, we will explain the multilingual design and illustrate the different ways in which equivalence relations can be accessed in the database. As we will see, the mappings across the wordnets may result in a more loose pattern of lexicalizations around a concept: abstracting from the language-specific lexicalization in the source wordnet. The hierarchical structure in the target wordnet can then be used to generate the correct lexical items or combinations of items, given the requirements of the translation context. In this respect, the conceptual mapping in EuroWordNet resembles 'Shake and Bake methods' in MT (Whitelock 1992). In section 3, we will describe how the equivalence relations have been extracted for the Dutch wordnet.

2. The multilingual design of the database[2]

In the EuroWordNet database, the wordnets are stored as independent modules with language-internal relations between synsets. Connections between these synsets are established via a separate equivalence link to the Inter-Lingual-Index (ILI). Each synset in the monolingual wordnets has at least one equivalence relation with a record in this ILI. Language-specific synsets linked to the same ILI-record may be equivalent across the languages. This is illustrated in Figure 1, which is taken from the graphical interface to the EuroWordNet database, called Periscope (Cuypers and Adriaens 1997). The top-half of the screen-dump shows a window with a fragment of the Dutch wordnet at the left and a similar fragment of WordNet1.5 at the right. The bottom window shows a similar parallel view for the Italian and Spanish wordnets. Each synset in these windows is represented by a rectangular box followed by the synset members. On the next line, the closest Inter-Lingual-Index concept is given, following the = sign (which indicates direct equivalence). In this view, the ILI-records are represented by an English gloss. Below a synset-ILI pair, the language-internal relations can be expanded, as is done here for the hyperonyms. The target of each relation is again represented as a synset with the nearest ILI-equivalent (if present). The first line of each wordnet gives the equivalent of *cello* in the 4 wordnets. In this case, they are all linked to the same ILI-record, which indirectly suggests that they should be equivalent across the wordnets as well. We also see that the hyperonyms of *cello* are also equivalent in the two windows, as is indicated by the lines connecting the ILI-records. Apparently, the structures are parallel across the Dutch wordnet and WordNet1.5 on the one hand and the Spanish and Italian wordnets on the other. However, we see that the intermediate levels for *bowed stringed instrument* and *stringed instrument* in the Dutch wordnet and WordNet1.5 are missing both in Italian and Spanish. Had we compared other wordnet pairs, the intermediate synsets would be unmatched across the wordnets.

The advantages of an interlingua such as the ILI are well-known in MT translation (Copeland et al. 1991, Nirenburg 1989):

1. it is not necessary to specify many-to-many equivalence relations between each language-pair and to have consensus across all the groups on the equivalence relations: each group only considers the equivalence relations to the Index.
2. new languages can be added without having to reconsider the equivalence relations for the other languages.
3. it is possible to adapt the Inter-Lingual-Index as a central resource to make the matching more efficient or precise.



Query | Nederlands | cello | WordNet1.5

violoncel :1 cel :5 cello :1
= large; seated player holds it upright wh
has hyperonym
strijkinstrument :1
= "the strings played superlative
has hyperonym
snaarinstrument :1
= a musical instrument
has hyperonym

violoncello :1 cello :1
= large; seated player holds it upright while playir
has hyperonym
string :3
bowed stringed instrument :1
= "the strings played superlatively well"
has hyperonym
stringed instrument :1
= a musical instrument in which

Query | Italiano | violincello | Español | cello

violoncello :1
= large; seated player holds it upright while play
has hyperonym
strumento :4
strumento musicale :20
= an instrument used to produce music

chelo :1 violoncelista :1 violoncelo :1
violonchelo :1 cello :1
= large; seated player holds it upright while play
has hyperonym
instrumento_musical :1
instrumento :4
= an instrument used to produce music

For Help, press F1 | Ln 0 LnCnt 5 | Filter On

Figure 1: Parallel wordnet structures in EuroWordNet linked to the same ILI-records.

Another important feature is that the ILI is an unstructured list of concepts with the only purpose of linking word meanings across languages. Each concept is solely represented by a so-called ILI-record containing a synset and a gloss. No semantic relations, such as hyponymy or meronymy, are expressed between these records. As an unstructured fund of concepts, there is no need to agree on a universal ontology shared by all the languages, and it will be easier to adapt the index when new languages are added. [3] In (Vossen et al. 1997: 1-8) and (Peters et al. 1998: 221-251) further details are given on the multilingual design and functionality of the database.

Because the ILI is unstructured, different parallelisms and structural mismatches can be easily expressed, without complicating the comparison. This not only holds for the language-internal relations but also for the equivalence relations. In the above example, the synsets are related to the ILI by a direct equivalence relation. In the next subsection we will see that there can be also complex equivalence relations with ILI-records. Complexity and efficiency of the matching is on the other hand reduced by adapting the index itself. Initially, the ILI consists of WordNet1.5 synsets but, along the project, the ILI is extended (see Peters et al, 1998a, Peters et al. 1998b: 221-251). This will be discussed in subsection 2.2. Subsection 2.3. then gives an overview of the most important matching possibilities.

2.1. Complex equivalence relations

Parallel to the language-internal relations (see Alonge et al., 1998: 91-115) there are six different types of inter-lingual relations. The most straight forward relation is EQ_SYNONYM, which applies to meanings which are directly equivalent to some ILI-record, as has been shown in Figure 1 above. In addition there are relations for complex-equivalence relations, among which the most important are:

- EQ_NEAR_SYNONYM when a meaning matches multiple ILI-records simultaneously, or when multiple synsets match with the same ILI-record.
- EQ_HAS_HYPERONYM when a meaning is more specific than any available ILI-record.
- EQ_HAS_HYPONYM when a meaning can only be linked to more specific ILI-records.

The complex-equivalence relations are comparable to the kinds of mismatches across word meanings described in the Acquilex project in the form of complex *TLINKS* (Ageno et al 1993, Copestake et al. 1995, and Copestake and Sanfilippo 1993). It is possible to manually encode these relations directly in the database, but they can also be extracted semi-automatically using the technology developed in Acquilex. The difference between Acquilex and EuroWordNet is that the *TLINKS* in Acquilex are lexical transfer links between language-pairs at a sense-level, whereas the equivalence relations in EuroWordNet are established at the synset level from each language to a single interlingua (the ILI). Language-to-language mappings can only indirectly be inferred via the ILI.

In EuroWordNet, the complex relations are needed to help the relation assignment during the development process when there is a lexical gap in one language or when meanings do not exactly fit. The first situation, in which a single synset matches several ILI-records simultaneously, occurs quite often. The main reason for this is that the sense-differentiation in WordNet1.5 is more fine-grained than in the traditional resources from which the other wordnets are built. For example, in the Dutch resource there is only one sense for *schoonmaken* (to clean) which simultaneously matches with at least 4 senses of *clean* in WordNet1.5:

- {make clean by removing dirt, filth, or unwanted substances from}
- {remove unwanted substances from, such as feathers or pits, as of chickens or fruit}
- {remove in making clean; "Clean the spots off the rug"}
- {remove unwanted substances from - (as in chemistry)}

The Dutch synset *schoonmaken* will thus be linked with an EQ_NEAR_SYNONYM relation to all these senses of *clean*. A similar situation may arise when there is under-differentiation in the Dutch wordnet. For example, *keuze* in the Dutch resource is defined as the *act* or *result* of choosing, likewise it can be linked with EQ_NEAR_SYNONYM relations to both *choice*#1 (the act of choosing) and *choice*#2 (what is chosen) in WordNet 1.5.

Despite the sense-differentiation in WordNet1.5, the reverse situation also occurs. For example, *versiersel* and *versiering* are not coded as synonyms in the Dutch resource but they can still both be linked to the same WN1.5 synset *decoration*. It may be the case that the Dutch words should be merged into a single synset, but, they can also be related by a weaker NEAR_SYNONYM relation. In the latter case, they can share the same ILI-record but the equivalence relation should be EQ_NEAR_SYNONYM and not EQ_SYNONYM.

The EQ_HAS_HYPERONYM is typically used for gaps in WordNet1.5 or in English. Such gaps can be cultural or pragmatic. A cultural gap is a concept not **known** in the English/American culture, e.g. the Dutch noun *citroenjenever*, which is a kind of gin made out of lemon skin, or the Dutch verb: *klunen* (to walk on skates over land from one frozen water to another). Pragmatic gaps are caused by lexicalization differences between languages, in the sense that in this case the concept is known but not expressed by a single lexicalized form in English., e.g.:

Dutch:	*doodschoppen* (to kick to death),
Spanish:	*alevín* (young fish),
Italian:	*rincasare* (to go back home).

In these cases the lexicalization patterns in the languages are different from English but the concepts are familiar to all cultures.

In all the above cases, the non-English word is more specific and thus can be related to a more general English ILI-concept using an EQ_HAS_HYPERONYM relation. The EQ_HAS_HYPONYM is then used for the reversed situation, when wordnet1.5 only provides more narrow terms. An example is Spanish *dedo* which can be used to refer to both *finger* and *toe*. In this case there can only be a pragmatic difference, not a genuine cultural gap.

A special case of gaps are mismatches in Part of Speech across languages, e.g. in Dutch the adjective *aardig* is equivalent to the verb *to like* in English but there is no verb with that meaning in Dutch. The equivalence relations to the ILI are however not sensitive to the Part-of-Speech. It is thus possible to express an EQ_NEAR_SYNONYM relation between *aardig* Adjective and *like* Verb.

The complex equivalence relations are expressed separately from each language to the index. Decisions on the matching are taken by each site separately for their language, towards the English ILI. In addition, there is also an effort to smoothen the matching across the wordnets by adapting the index. This will be discussed in the next subsection.

2.2. Adapting the Inter-Lingual-Index

There are two kinds of adaptations to the ILI to improve the matching across the wordnets:

1. if a language-specific meaning is missing in WordNet1.5 a new index item (properly glossed in English) is added to represent it.
2. word meanings in Wordnet1.5 that show regular polysemy or extreme sense-differentiation are grouped by globalized meanings in the form of so-called Composite ILI-records

The first measure is necessary to express a matching between synsets in cases where there is no equivalent in English (or WordNet1.5) but there is a matching between two other languages. The ILI should thus be the superset of all concepts occurring in the wordnets. Gaps, as discussed above, have to be filled following a coordinated updating procedure. All sites will send descriptions of the gaps in the form of potential new ILI-records to one site. The ILI-records will be described using a formalized semantic specification so that the candidates can be compared. If there is sufficient overlap between at least two descriptions, a new ILI-record is added and the local synsets referring to this new ILI-record will get an additional EQ_SYNONYM relation to this record. These synsets will thus have at least two different equivalence relations, a complex equivalence relation to the closest WordNet1.5 synset and a simple equivalence relation to the new ILI-record, e.g.:

Spanish Wordnet		**ILI**		**Italian Wordnet**
dedo	eq_hyponym	{*toe}*	eq_hyponym	dito
	eq_hyponym	{*finger}*	eq_hyponym	
	eq_synonym	{*finger or toe*}	eq_synonym	

This example shows that it is possible to extract direct equivalences in Italian and Spanish, but also to find the closest matches with English (albeit more specific).

Even though the ILI should be the superset of concepts occurring in the different wordnets, it should, on the other hand, not be too fine-grained either. If many subtle senses are distinguished, it is much more complicated to establish equivalences across the wordnets. In the case of "clean", for example, it may be that different sites link equivalent synsets to different meanings, resulting in a mismatch across the languages. A similar mismatch may be caused by inconsistent enumeration of regular polysemy across resources. In the ILI, there are different synsets for *university* as a building and *university* as the organization, and in fact many institute/building pairs are present. However, in other wordnets we may find situations where only one of the senses is given. If a different choice is made for the *building* or the *institute*, synsets can not be matched across wordnets.

The second adaptation to the ILI therefore aims at grouping senses that can be related by 'regular polysemy' (Apresjan 1973; Copestake and Briscoe 1991; Nunberg and Zaenen 1992). This is achieved by adding so-called Composite ILI-records, which can be compared with Complex Types as defined by Pustejovsky (1995). The next example shows such a Composite ILI-record for "university" that relates *university* 1 (the BUILDING) and *university* 2 (the INSTITUTE), where the third row specifies the polysemy relation that holds between them.

ILI-ID	@62489@
Synset	university
Polysemy-type	Metonymy
Source-references	
Source-id	2039764
Word form	university
POS	NOUN
Original-ILI-ID	@12547@
Gloss	(where a seat of higher learning is housed, including administrative and living quarters as well as facilities for research and teaching)
Source-id	5276749
Word form	university
POS	NOUN
Original-ILI-ID	@35629@
Gloss	(the faculty and students of a university);

Whenever such a Composite ILI-record is added to the ILI, the EuroWordNet database will automatically generate additional equivalence relations for all synsets in the wordnets related with an EQ_SYNONYM or EQ_NEAR_SYNONYM relation to any of the specific meanings that are grouped by this ILI-record. So if Dutch, Italian and Spanish equivalents of *university* are linked to the *building*, *institute* or both, they will now get an additional EQ_METONYM relation to the Composite ILI-record for *university*:

Dutch	**Italian**	**Spanish**
universiteit 1 {institution}	universitare 1 {building}	universidad 1 {institution; building}
ILI-reference	ILI-reference	ILI-reference
Eq_synonym @35629@	Eq_synonym @12547@	Eq_near_synony @12547@
Eq_metonym @62489@	**Eq_metonym @62489@**	Eq_near_synonym @35629@
universiteit 2 {building}		**Eq_metonym @62489@**
ILI-reference		
Eq_synonym @12547@		
Eq_metonym @62489@		

Note that it is not necessary that the metonymy-relation also holds in the local language. In this example only the Dutch wordnet has two senses that parallel the metonymy-relation in the ILI. The relation between these two Dutch senses is now also encoded via the metonymy-equivalence relation with the more global ILI-record. The Italian and Spanish example only list one sense (which may be correct or an omission in their resources). In the case of Spanish there are multiple equivalences of both senses of *university*, whereas the Italian synset is only linked to the *building* sense. The Spanish example is, in fact, equivalent to the new globalized ILI-record. Even though none of the local wordnets has the same differentiation, all four meanings now share the metonymy link and, likewise, can be retrieved in a global way when we look for synsets to the same ILI-record with EQ_METONYM.

Similar Composite ILI-records are added for generalizations that group over-differentiation as we have seen for "clean" (related by EQ_GENERALIZATION) and for enumerated senses that reflect diathesis alternations for verbs (related by EQ_DIATHESIS), such as between causative and inchoative pairs, e.g.:

hit 1:	hit a ball (synonym: cause to move by striking)
hit 2:	come into sudden contact with: “The arrow hit the target”
hit 3:	deal a blow to; “He hit her hard in the face”

Differences in arity and the semantic characterization of subcategorized arguments highlight different perspectives on the situation described by the predications, or express semantic notions such as ‘causation’ and ‘result of causation’ (Levin 1993). By relating these diathesis alternation patterns to more Composite ILI-records we will thus be able to link local synsets regardless of whether the verbs in question display dissimilar alternation patterns in different senses, have a number of alternations collapsed in a single sense, or are monosemous. Buitelaar (1998) and Peters et al. (1998a) describe how these sense-groups can be extracted from a resource such as WordNet1.5.

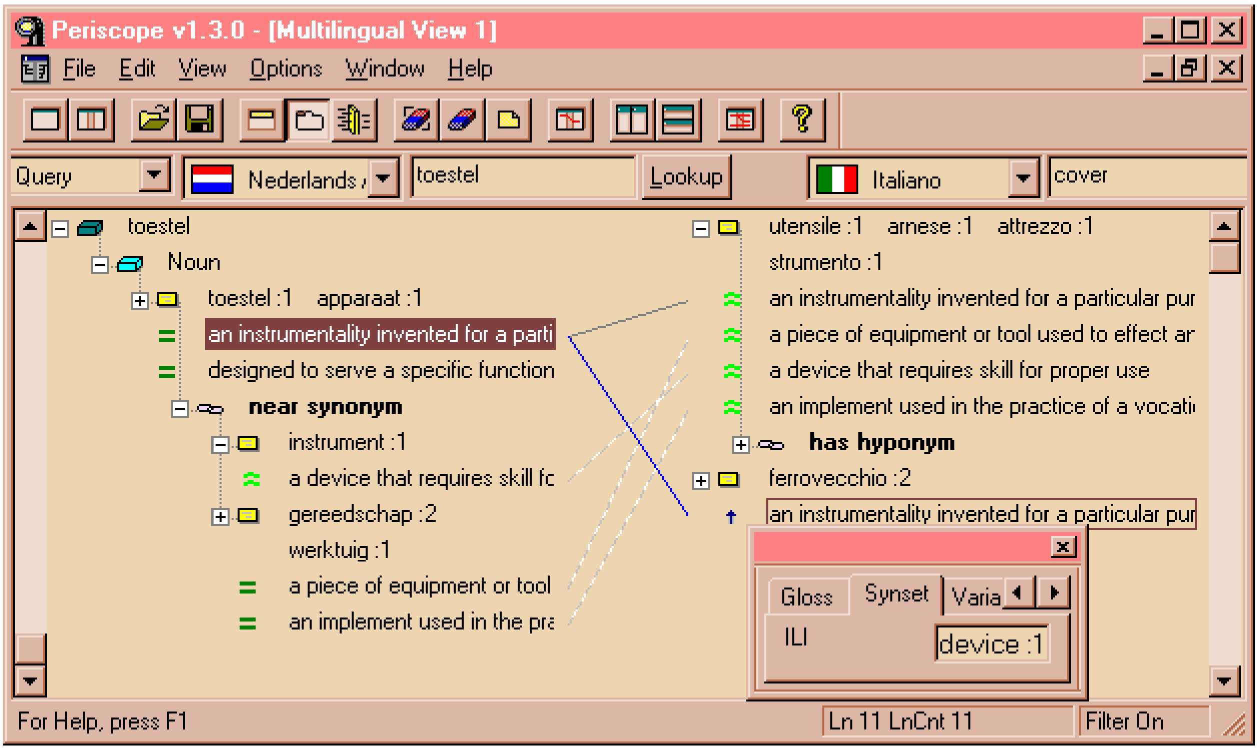
2.3. Accessing complex equivalence mappings

From what has been said so far it follows that there can be many-to-many mappings from local synsets to ILI-records. This may either be an EQ_NEAR_SYNONYM relation from and/or to multiple synsets (possibly with different part-of-speech), or an EQ_HAS_HYPONYM/ EQ_HAS_HYPERONYM and an EQ_SYNONYM to a new ILI-record, or various combinations of these (or other types of equivalence relations). Finally, it is possible that a single synset in a wordnet is linked to both a Composite ILI-record with an EQ_METONYM, EQ_DIATHESIS or EQ_GENERALIZATION and to one of the more specific senses grouped by the Composite ILI.

Table 1 : Overview of mapping relations to the ILI

Relation	*POS*	*Source Synsets : Target ILIs*	*Example*
eq_synonym	same	1:1	auto : car
eq_near_synonym	any	many : many	apparaat, machine, toestel : apparatus, machine, device
eq_hyperonym	same	many : 1 (usually)	citroenjenever: gin
eq_hyponym	same	(usually) 1 : many	dedo : toe, finger
eq_metonymy	same	many/1 : 1	universiteit, universiteitsgebouw: university
eq_diathesis	same	many/1 : 1	raken (cause), raken: hit
eq_generalization	same	many/1 : 1	schoonmaken : clean

Note that a many-to-many mapping from a wordnet to the ILI, may also cause a further spreading when multiple ILI-records are next mapped to another wordnet. In the next screen-dump we see how such a fuzzy mapping results for *machine*, *apparatus*, *tool* in Dutch and Italian. In this example, 3 near synonyms in the Dutch wordnet are linked to multiple ILI-records, from-top-to-bottom: *device*, *apparatus*, *instrument*, *implement*, *tool*. The ILI-records are again represented by their glosses, where the synset of the highlighted ILI-record (device:1) is shown in the small box at the bottom-right corner. In the Italian wordnet we see that 4 of these ILI-records are given as EQ_NEAR_SYNONYMs of a single synset *utensile:1* but *device* is linked to *ferrovecchio:2* by an EQ_HAS_HYPERONYM relation (as indicated by the symbols).



toestel
Noun
toestel :1 apparaat :1
= an instrumentality invented for a parti
= designed to serve a specific function
near synonym
instrument :1
≈ a device that requires skill fc
gereedschap :2
werktuig :1
= a piece of equipment or tool
= an implement used in the prε

utensile :1 arnese :1 attrezzo :1
strumento :1
≈ an instrumentality invented for a particular pur
≈ a piece of equipment or tool used to effect ar
≈ a device that requires skill for proper use
≈ an implement used in the practice of a vocati
has hyponym
ferrovecchio :2
† an instrumentality invented for a particular pur

ILI device :1

Figure 2: Many-to-many mappings of near synonyms of *apparatus* synsets to ILI-records.

Another important characteristic of the equivalence relations is the fact that they are established at the synset level. This is different from a traditional bilingual dictionary where specific relations are expressed between individual words or word-senses. For example, a pejorative term such as "idiot" is usually translated in a bilingual dictionary by a pejorative term in a the target language. In EuroWordNet, both the pejorative and the neutral term are members of the same synset and may have a single ILI-record as equivalent.

We can thus say that, in general, the effect of the multilingual relations in EuroWordNet is that concepts are matched rather than words, that multiple concepts may share ILI-records (index-terms) or single concepts may yield multiple ILI-records. Furthermore, the ILI may be accessed very specifically by EQ_SYNONYM relations only, or by indicating any of the other complex equivalence mappings. The database thus provides the possibility to project a single concept or a cluster of concepts to another language, either specifically or in a more fuzzy way.

Once we have accessed a cluster of concepts in the target language, we can further use the language-internal relations to see the conceptual dependencies between these words (and possibly other words). This may point to solutions for gaps in the target language as is illustrated in Figure 3, where Dutch compound verbs for *ways of killing* are not lexicalized in English.

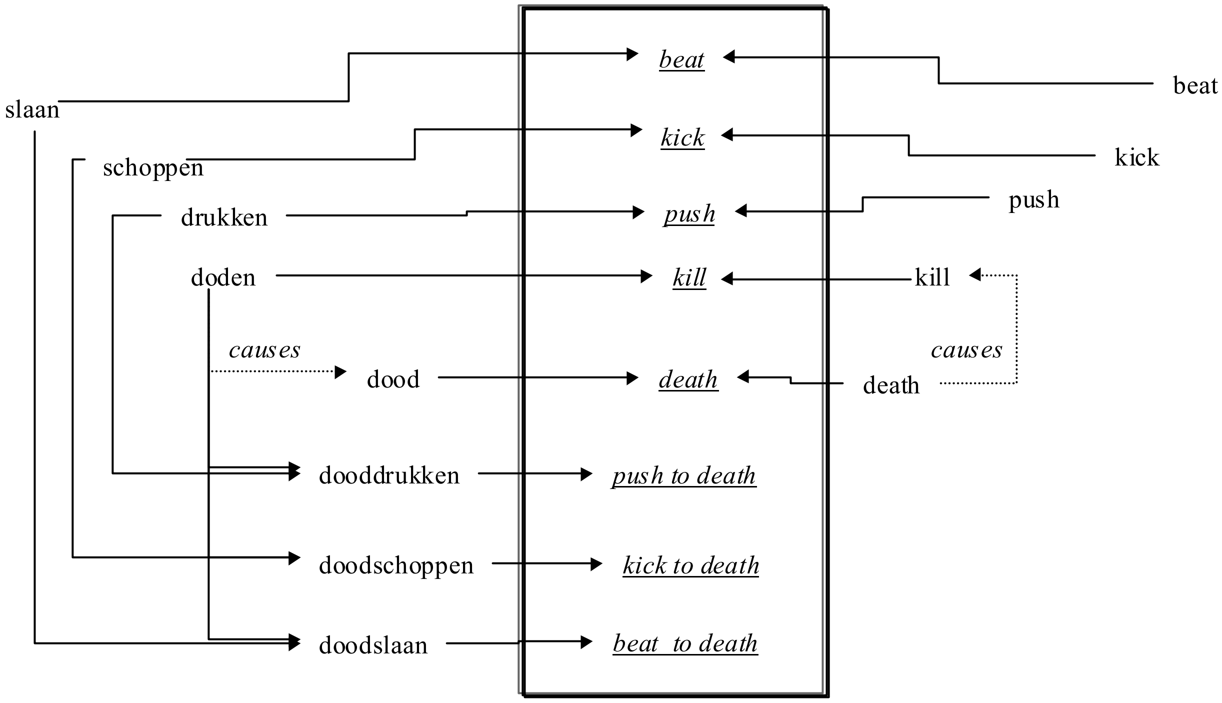


Figure 3: *Ways of killing* lexicalized in Dutch and not in English.

Here we see that the ILI is extended to represent concepts for the Dutch verbs, and there is no mapping to English verbs at the right side. The Dutch verbs have multiple hyperonyms to both the manner in which the event takes place (*beat*, *kick*, *push*) and the result (*kill*). Furthermore, *doden* and *kill*, which are equivalents, have a causal relation to the nouns *dood* and *death*, which are equivalent too. From this we may develop a strategy to generate expressions such as "kill by kicking" or "kick to death" as equivalents for the Dutch verb "doodschoppen".

Concluding, we can say that instead of a single or a few specific alternatives in a bilingual dictionary, the EuroWordNet database gives a more comprehensive overview of concept-lexicalization in the target language, from which to choose the best candidate. In this sense, we can make a parallel with the 'Shake and Bake' methodology in Machine Translation (Whitelock 1992), where first an abstraction is made from the structural properties in the Source Language to a more neutral conceptual level (Shake), and next a (possibly different) new structure is generated in the target language (Bake). In the case of EuroWordNet, we are dealing with lexical Shake: abstract from the lexicalization that may be specific for a language. Bake is then possible by selecting the most appropriate candidate on the basis of co-occurrence restrictions in the target language, or the pragmatic and morpho-syntactic properties of the members in the synset. This kind of information can be extracted from Parole lexicons properly linked to the EuroWordNet database.

3. Establishing equivalence relations

The extraction of the equivalence links between the Dutch wordnet and WordNet1.5 is partly done by hand and partly using automatic techniques. The manual coding is carried out for the most important concepts in the database and for those concepts that have been poorly matched by the automatic techniques. Important concepts are those synsets which have many meanings in the wordnets and/or high positions in the hierarchy (the so-called Base Concepts, Rodriguez et al. 1998: 117-152). The manual encoding of the most important concepts had to ensure that the cores of the wordnets are well-matched. In total, 7,521 synsets have been translated by hand (4,138 nominal 3,383 verbal synsets). These figures include the manual encoding of poor translations, which will be further discussed below.

The remaining synsets have been translated by mapping the Van Dale database with the bilingual Dutch-English dictionary and mapping the translations to WordNet1.5. Note that such a mapping is carried out for synsets. It is therefore sufficient if one member of a synset can be mapped. Nevertheless, a proportion of the original Dutch database did not receive a translation, either because the entry was missing in the bilingual dictionary, or the translations could not be found in WordNet1.5. In these cases we tried to directly look up the Dutch word in WordNet1.5 (loan words from English have been omitted in the Dutch-English dictionary) or we tried to translate it by looking it up in the reversed English-Dutch dictionary (where we created entries for the Dutch translations with the original English entries as translations). The result of this procedure is shown in the next table:

Table 2: Synsets without translation in the Dutch database

	Number of Synsets in the Dutch Database	*Number of Synsets without translation to WordNet1.5*	*%*
nouns	52359	20511	39,17%
verbs	9125	1060	11,62%

The result for verbs (11,62% not translated)are much better than for nouns (39,17% not translated). This is due to the fact that the nominal part contains more specialized vocabulary.

Once a matching entry has been found in WordNet1.5, all the senses of the entry are proposed as possible translations. If there is only one synset translation the procedure stops, and we assume that this translation is correct. If there are multiple translations they are weighted by measuring the conceptual-distance in WordNet1.5. The conceptual distance measurement is based on Agirre and Rigau (1996) who calculate the distance between senses by counting the steps to their closest shared node in the network, taking into account the level of the hierarchy and the density of nodes relative to the average density. There are two situations for which the conceptual distance is calculated:

- the distance between senses of multiple alternative translations of a single entry in the bilingual dictionary.
- the distance between each possible translation and the translations of hyponyms and hyperonyms of the Dutch word

The first situation occurs when, for example, the Dutch word *orgel* has two translations, *organ* and *keyboard*, for the same sense. Since the polysemy of these translations is often not parallel, it is possible to favor the sense of *organ* and *keyboard* that have the shortest distance. The second situation is illustrated in Figure 4. Here we see that *orgel* in Dutch is translated as *organ,* which can either be a *musical instrument* or a *body part*. Since the hyperonym and a hyponym of *orgel* in the Dutch wordnet have already been translated it is possible to measure the distance of the two senses of *organ* to the translations of the hyperonym and hyponym:

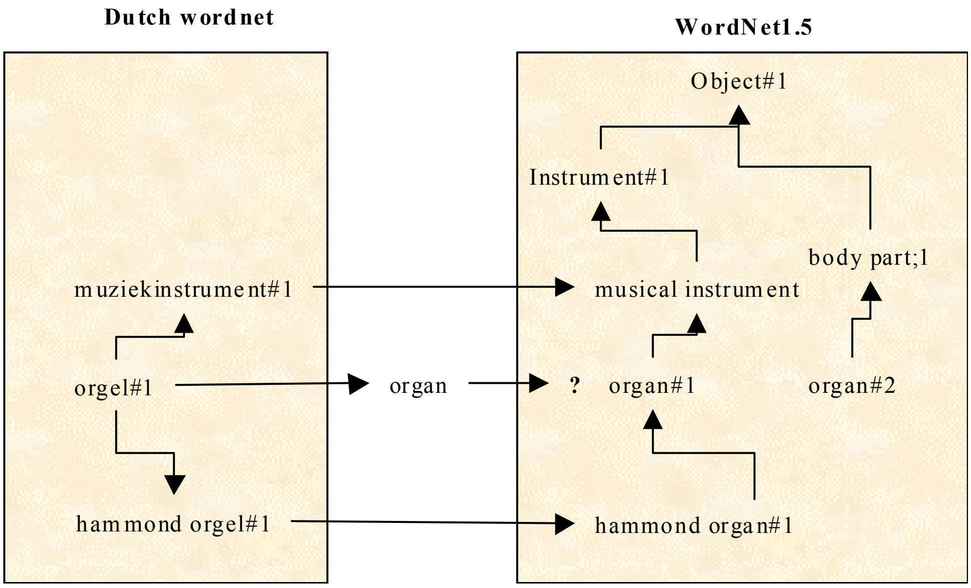


Figure 4: Selecting translations to WordNet1.5 by distance to the translated context in the Dutch wordnet.

The distance measuring of the translations to the context in the Dutch wordnet, leads to a ranking of all the senses of a translation. The heuristics for automatically deriving equivalence relations are implemented in such a way that bad matches are removed if the best match is above a certain threshold. If not, all matches are maintained. A large number of matches with a low score therefore indicates that the system had poor evidence for matching. Furthermore, if the best match is above a threshold, all matches below a specified percentage of the best match (e.g. less than 70% of the best score) are removed. The implication of this is that the heuristics will remove poor matches when there is a strong differentiation in matching but will tend to keep matches when they are relatively close. By searching the database for synsets which have an extremely high number of automatically-derived translations (with a relatively low score) we can isolate dubious cases. This is shown for the next example "inlassen" (to weld something in between something else), where none of the suggested translations is correct (in fact, the correct translation probably does not exist):

Dutch synset: inlassen \# 2

2.07	00604079-v	bring out#3; introduce#6
1.77	00121079-v	alter#4; falsify#1; interpolate#1
1.7	00579406-v	insert#5; slip in#1; sneak in#1; stick in#2
1.65	00437968-v	barge in#1; break in#4; butt in#1; chime in#1; cut in#4; put in#2
1.65	00514811-v	come in#2; inject#3; interject#1; interpose#1; put in#3; throw in#1
1.35	00361286-v	extrapolate#2; interpolate#2
1.30	00818159-v	enter#3; infix#3; insert#7; introduce#7
1.27	01417019-v	admit#4; allow in#1; let in#2
1.13	00507610-v	introduce#5; preface#2; premise#3
1.00	00799930-v	insert#6; tuck#3
0.946	00927659-v	introduce#8
0.946	00939471-v	innovate#1; introduce#9
0.914	00507320-v	acquaint#2; introduce#4; present#6
0.898	00397690-v	introduce#3
0.668	00210341-v	inaugurate#1; introduce#2; usher in#1
0.668	01189328-v	bring in#2; introduce#10
0.668	01297479-v	hive away#1; lay in#1; put in#5; salt away#1; stack away#1; stash away#1; store#7
0.668	01532350-v	put in#6
0.544	01386819-v	admit#3; include#3; let in#1; let participate#1
0.364	00113224-v	insert#4; introduce#1; put in#1; stick in#1
0.243	00605466-v	put in#4; submit#5

The number-codes, such as 00605466-v and 00113224-v, are file-offset positions that uniquely identify a synset in WordNet1.5. In general, we can thus state that many matches or low scores indicate poor matches. We have therefore manually translated all verbal synsets with more than 20 translations and all nominal synsets with more than 30 translations. Next we looked at:

- polysemous words with many meanings and many translations
- synsets with many relations and many translations

In many cases, it appeared that polysemous words with a badly translated sense often had poor translations for the other senses as well. We then manually translated all the senses of such a polysemous words. In addition, we have looked at words with many relations and many translations. All verbs with more than 2 relations and more than 10 translations have been manually translated as well. The same holds for nouns with more than 10 relations and more than 10 translations. About 3,000 synsets with low quality matches have thus been translated by hand.

Since many translations have been improved manually, we expected an improvement of the tree-matching effect for synsets related to these concepts (by hyponymy or hyperonymy) as well. After a manual revision of worsed cases we ran the tree-matching option again. The Table 4a and 4b below show the results. In addition we have also applied 2 other heuristics:

- reversing possible translations via an English-Dutch dictionary.
- matching the overlap of top-concepts.

The bilingual dictionaries from Van Dale are intended for Dutch speakers and, therefore, are not bi-directional. We therefore expect that the vocabularies and translations in the Dutch-English and English-Dutch dictionary are not the same. If several synset members in WordNet1.5. have the same Dutch word as a translation in the English-Dutch dictionary, or several Dutch translations which are in the same Dutch synset, then this can be seen as additional evidence for the correctness of a translation. We have thus created a separate bilingual database from the English-Dutch dictionary of Van Dale (Martin and Tops 1989). This resource is then used to see which translations are reversible, according to the following algorithm:

```
1. take the possible candidate translations generated from the Dutch-English resource to WordNet1.5.
2. look up the target variants in the English-Dutch resource
3. increase the match:
    3.1. each time an English variant has the Dutch source as its translation
    3.2. if multiple Dutch variants are given as the translation for a single English sense
```

The next example illustrates this for the Dutch synset "lakken" (coat with lacquer):[4]

```
Dutch Variants: lakken;
WordNet Match: 00779724-v
        WordNet Variant: affix a seal to
        WordNet Variant: seal
        WordNet Variant Translation:  op robben/zeehondenvangst gaan/zijn;
        Overlap = 0
WordNet Match: 00726098-v
        WordNet Variant: coat with lacquer
        WordNet Variant: lacquer
        WordNet Variant Translation:  lakken;
        WordNet Variant Translation:  vernissen;
        Overlap = 1
        From: 26.6  To: 39.8
```

The verb "lakken" has a translation candidate synset 00779724-v. The first variant of this synset is "affix a seal to". This cannot be found in the English Dutch dictionary. The second variant "seal" can be found. However, the translation of "seal" does not contain the original Dutch word "lakken" (the overlap between the translations and the original synset members is 0). The next WordNet Match is the synset 00726098-v. The first synset member "coat with lacquer" cannot be found but the second "lacquer" is found and has "lakken" as one of its translations. The matching for this synset will thus be increased (From: 26.6 To: 39.8).

The second heuristics makes use of the fact that we have separately added 63 Top Concepts (TCs) to the Base Concepts (the most important concepts) in the Dutch wordnet and WordNet1.5. The TCs represent fundamental semantic features, such as Natural, Artifact, Dynamic, Static, Physical, Mental that can be combined into complex feature combinations (see Rodriguez et al. 1998: 117-152, for a further motivation and explanation of the top-ontology). By inheriting these TCs to more specific concepts via the hyponymy relations it is possible to measure the overlap in TCs between Dutch senses and their candidate translations. If a candidate translation has many overlapping TCs, it is a more likely candidate for translating. In the next example, the Dutch word "hart" inherits the top-concepts *Living* and *Part* from its hyperonyms, which it shares only with sense 4 of the senses of "heart" in WordNet1.5:

hart 1
orgaan 1 (*Living Part*) deel 2 (*Part*) iets 1 LEAF

---

heart 1
playing card 1 card 1 (*Artifact Function Object*) paper 6 (*Artifact Solid*)
material 5 (*Substance)* matter 1 inanimate object 1 entity 1 LEAF
heart 2
disposition 2 (*Dynamic Experience Mental*) nature 1
trait 1 (*Property*) attribute 1 (*Property*) abstraction 1 LEAF
heart 3
bravery 1 spirit 1 character 1 trait 1 (*Property*) attribute 1 (*Property*)
abstraction 1 LEAF
heart 4
internal organ 1 organ 4 (*Living Part*) body part 1 (*Living Part*)
part 10 entity 1 LEAF

This heuristics is expected to be especially useful for discriminating translations of verbs because their semantics is less dependent on the hierarchical structure (which is relatively flat and shallow). A rich encoding with features for verbs with a poor hyponymic structure can still contain sufficient evidence for choosing translations. The effect of this matching obviously depends on the coverage of the features and the diversity of features. Currently, we have limited ourselves to the 63 features from the EuroWordNet top ontology. To improve the matching it is possible to add more discriminative features at crucial points of both hierarchies. To get a maximal coverage of inherited top-concepts we ensured that all tops in the Dutch wordnet and in WordNet1.5 are classified according to the ontology, and that most tops in the Dutch wordnet are unified into a minimal number of trees. For WordNet1.5. we had to add TCs to 389 verbal synsets and 2 nominal synsets, which are tops but have not previously been classified by the TCs, In total 2006 Dutch synsets (1170 nouns and 836 verbs) and 1410 WordNet1.5 synsets (793 nouns and 617 verbs) have been classified with one or more top-concept features. Furthermore, we have converted the lexicographer's file codes in WordNet1.5 to compatible EuroWordNet top-ontology codes, as is indicated in the next table. Since all synsets in WordNet1.5 have been assigned by these codes we thus get a very high coverage of the semantic features.

Table 3: Conversion of WordNet1.5 Lexicographer's file codes to EuroWordNet top-concepts

Code	WordNet File Name	EuroWordNet Top Concepts
04	noun.act	Agentive;
05	noun.animal	Animal;
06	noun.artifact	Artifact;
07	noun.attribute	Property;
08	noun.body	Object; Natural;
09	noun.cognition	Mental;
10	noun.communication	Communication;
11	noun.event	Dynamic;
12	noun.feeling	Experience;
13	noun.food	Comestible;
14	noun.group	Group;
15	noun.location	Place;
16	noun.motive	3rdOrderEntity;
17	noun.object	Object;
18	noun.person	Human;
19	noun.phenomenon	Phenomenal;
20	noun.plant	Plant;
21	noun.possession	Possession;
22	noun.process	Dynamic;
23	noun.quantity	Quantity;
24	noun.relation	Relation;
25	noun.shape	Physical;
26	noun.state	Static;
27	noun.substance	Substance;
28	noun.time	Time;
29	verb.body	Dynamic; Physical;
30	verb.change	Dynamic;
31	verb.cognition	Mental; Dynamic;
32	verb.communication	Communication; Dynamic;
33	verb.competition	Social; Dynamic;
34	verb.consumption	Physical;Usage; Dynamic;
35	verb.contact	Location; Physical
36	verb.creation	Existence; BoundedEvent;
37	verb.emotion	Experience; Mental;
38	verb.motion	Location; Physical; Dynamic;
39	verb.perception	Experience; Physical; Dynamic;
40	verb.possession	Possession; Dynamic;
41	verb.social	Social; Dynamic;
42	verb.stative	Static;
43	verb.weather	Phenomenal; Physical; Dynamic;

The effects of the above measures are shown in the table 4a and 4b below. We took a random sample of nouns and verbs and measured the quality of the matching by scoring how often the highest match was correct, the 2nd highest match, etc.. This has been done for the Dutch wordnet first with minimal manual encoding of equivalence relations, and next after taking each of the above measures in sequence to each result: 1) encoding dubious translations and important synsets by hand and after that running the tree-matching algorithm again, 2) applying the reverse translation option using the English-Dutch dictionary, 3) applying the top-concept matching. In the table, the rows indicate rank of the correct match: the first row the number of times the highest match was correct (match 1), the 2nd correct, the 3rd, 4th, 5th and higher. Obviously, in the ideal case the highest match (rank 1) should be correct. The next row indicates the number of synsets that cannot be translated (presumably a gap in English), which can be translated but there correct translation was not present (non ok) or only a hyperonym translation is given (hyper). Finally, the number of synsets without a translation have been given. The columns then give the improvements, where the first column gives the figures and percentages wordnet with minimal manual encoding of equivalence relations (only the Base Concepts), the second column the results after the manual revision of dubious translations, the third column the results of making use of reversed translations and the fourth column the results of matching the top-concepts. The improvements are applied in a cascade. The final column gives the total gain with respect to the first column.

Table 4a: Automatic matching results for nouns

Match Rank	***Tree-matching before manual improvement***		***Tree-Matching after manual improvement***		***Reversed translation***		***Top-Concept Matching***		***Total Gain***
1	60	64,5%	70	66,6%	72	68,5%	74	70,4%	5,9%
2	8	8,6%	10	9,5%	8	7,6%	9	8,5%	-0,03%
3	4	4,3%	3	2,8%	3	2,8%	2	1,9%	-2,4%
4	0	0,0%	2	1,9%	3	2,8%	1	0,9%	0,9%
5	2	2,1%	2	1,9%	1	0,9%	1	0,9%	-1,2%
>5	1	1,0%	1	0,9%	2	1,9%	2	1,9%	0,8%
gap	6	6,4%	9	8,5%	10	9,5%	10	9,5%	3,0%
non ok	9	9,6%	3	2,8%	4	3,8%	4	3,8%	-5,8%
hyper	3	3,2%	5	4,7%	2	1,9%	2	1,9%	-1,3%
Subtot.	93		105		105		105		
notrans	102	52,3%	90		90		90		-6,1%
Total	195		195		195		195		
Top-3	72	77,4%	83	79,0%	83	79,0%	85	80,9%	3,5%

Table 4b: Automatic matching results for verbs

Match Rank	***Tree matching before manual improvement***		***Tree-Matching after manual improvement***		***Reversed translation***		***Top-Concept Matching***		***Total Gain***
1	28	33,3%	28	32,9%	32	37,6%	39	45,8%	12,5%
2	7	8,3%	14	16,4%	18	21,1%	12	14,1%	5,7%
3	9	10,7%	10	11,7%	4	4,7%	5	5,8%	-4,8%
4	2	2,3%	3	3,5%	4	4,7%	5	5,8%	3,5%
5	1	1,1%	4	4,7%	3	3,5%	4	4,7%	3,5%
>5	0	0,0%	6	7,0%	4	4,7%	2	2,3%	2,3%
gap	8	9,5%	10	11,7%	12	14,1%	10	11,7%	2,2%
non ok	19	22,6%	6	7,0%	4	4,7%	4	4,7%	-17,9%
hyper	4	4,7%	4	4,7%	4	4,7%	4	4,7%	-0,06%
Subtot	78		85		85		85		0,0%
notrans	6	7,1%	0		0		0		-7,1%
Total	84		85		85		85		
Top-3	44	52,3%	52	61,1%	54	63,5%	56	65,8%	13,5%

If we look at the first match (the highest matching score, 1) we see that for nouns each technique results in about 2% improvement. In total we gained 6% with respect to the first column. In the case of the verbs, we see that the tree-matching has not improved after the manual revision. This is expected because the general effect of tree-matching is poor for verbs. However, the reversed translation technique and, especially, the top-concept matching has resulted in a considerable improvement, 5% and 8% respectively. The total improvement for verbs is therefore even higher than for nouns: 12,55%. Obviously, an increase of correct first matches leads to a decrease of the lower matches. Note that, both for verbs and nouns, the number of gaps and the number of translated synsets has increased as well due to the fact that more word have been translated by the measure explained previously.

We applied a second count to measure the reliability of the matches in relation to the number of matches. Combined with the manually encoded equivalence relations this gives the following overall results:

Table 5: Quality of the equivalence relations

	Nouns			*Verbs*		
Matching Type	No of synsets	Perc.	Reliability	No of Synsets	Perc.	Reliability
manual/ok	4138	17,0%	100%	3383	37,0%	100%
1 match	4846	19,9%	86%	763	8,3%	78%
2 matches	3059	12,5%	68%	652	7,1%	71%
3-9 matches	5408	22,2%	65%	2471	27,0%	49%
10+ matches	1864	7,6%	54%	980	10,7%	23%
0 matches	5022	20,6%	n.a.	876	9,6%	n.a.
Total	24337			9125		

This table shows that automatic matches that give only 1 translation, but also with 2 translations, have a reasonable reliability (70-80% correct). This figure gets lower the more matches are left, where 10 or more translations are extremely unreliable. In many cases, we are then dealing with gaps which cannot properly be translated. Currently, the translations of these synsets are neglected and the synsets are automatically linked with an EQ_HYPERONYM relation to the translation of their hyperonym.

The final table then gives an overview of the equivalence relations that have been assigned so far. All automatically derived translations are of type EQ_NEAR_SYNONYM (although EQ_NEAR_SYNONYMs are occasionally also assigned by hand), the other types of equivalence relations are added manually:

Table 6: Equivalence Relations NL

Equivalence Relations	*Nouns*	*Verbs*	*Total*
EQ_BE_IN_STATE	14	2	16
EQ_HAS_HOLONYM	48	0	48
EQ_HAS_HYPERONYM	446	564	1010
EQ_HAS_HYPONYM	140	20	160
EQ_HAS_MERONYM	21	0	21
EQ_INVOLVED	2	13	15
EQ_IS_CAUSED_BY	3	15	18
EQ_NEAR_SYNONYM	28816	13190	42006
EQ_ROLE	9	0	9
EQ_SYNONYM	1730	275	2005
EQ_CAUSES	8	8	16
EQ_HAS_SUBEVENT	0	2	2
EQ_IS_SUBEVENT_OF	0	3	3
Total	31237	14092	45329

Because of the manual encoding of the unreliable manual translations, the number of EQ_HAS_HYPERONYM translations is relatively large among the complex equivalence relations. These are all (possible) gaps for which we could not find an appropriate translation by hand.

4. Conclusion

In this paper we explained the multilingual relations in EuroWordNet. We showed how complex and fuzzy equivalence relations are expressed, and how these can be used to get a global conceptual matching across wordnets that does not depend on the lexicalization in the source language: 'lexical shake and bake'. The wordnet structures can then be used to get a more comprehensive overview of the lexicalization in both the source and target language. Other information, such co-occurrence restrictions, morpho-syntactic properties, pragmatic features, should then be used to make the appropriate selection. Finally, we showed how the equivalence relations have been extracted using (semi)-automatic techniques. A large proportion of the Dutch wordnet (about 40%) has been translated automatically with a reliability ranging between 50-86%.

References

Ageno A., F. Ribas, G. Rigau, H. Rodriquez and F. Verdejo. 1993. *TGE: Tlinks Generation Environment*. Acquilex II (BRA 7315) Working Paper 7. Polytecnica de Catalunya, Barcelona.

Alonge, A., N. Calzolari, P. Vossen, L. Bloksma, I. Castellon, T. Marti, W. Peters. 1998. "The Linguistic Design of the EuroWordNet Database". In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds), *Special Issue on EuroWordNet. Computers and the Humanities*, Volume 32, Nos. 2-3 1998. 91-115.

Apresjan, J. 1973. "Regular Polysemy". *Linguistics* 142.

Buitelaar, P. 1998. *Corelex: Systematic Polysemy and Underspecification*, PhD., Department of Computer Science, Brandeis University.

Copeland, C., Durand, J., Krauwer, S. and Maegaard, B. (eds.). 1991. *The Eurotra Formal Specifications*, Office for Official Pblications of the European Community, Luxembourg.

Copestake A. and Briscoe, T. 1991. "Lexical operations in a unification-based framework". Ed. Pustejovsky J. and Bergler S. *Lexical Semantics and Knowledge Representation,* Association for Computational Linguistics.

Copestake A., T Briscoe, P. Vossen, A Ageno, I Castellon, F Ribas, G Rigau, H Rodriguez, A Sanmiotou. 1995. "Acquisition of Lexical Translation Relations from MRDs". *Journal of Machine Translation*, Volume 9, issue 3.

Copestake A. and A. Sanfilippo. 1993. *Multilingual Lexical Representation*, Acquilex II (BRA 7315) Working Paper 2. Cambridge University.

Copestake, A. 1995. "Representing Lexical Polysemy". *Proceedings of AAAI,* Stanford Spring Symposium, Stanford.

Cuypers, I. And G. Adriaens. 1997. *Periscope: the EWN Viewer*, EuroWordNet Project LE4003, Deliverable D008d012. University of Amsterdam, Amsterdam.

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

Díez Orzas, P., M. Louw and Ph. Forrest. 1996. *High level design of the EuroWordNet Database*. EuroWordNet Project LE2-4003, Deliverable D007.

Louw, M. 1998. *The Polaris User's Guide: The EuroWordNet Database Editor.* EuroWordNet (LE4-4003 Deliverable D024*)* University of Amsterdam..

Levin, B. 1993. English Verb Classes and Alternations, a Preliminary Investigation, University of Chicago Press, Chicago/London.

Martin W. and J. Tops (eds). 1989.*Groot woordenboek Engels-Nederlands.* Van Dale Lexicografie. Utrecht.

Martin W. and J. Tops (eds). 1986.*Groot woordenboek Nederlands-Engels.* Van Dale Lexicografie. Utrecht.

Miller . .A, R. Beckwidth, C. Fellbaum, D. Gross, and K.J. Miller. 1990. "Introduction to WordNet: An On-line Lexical Database".. *International Journal of Lexicography*, Vol 3, No.4, 235-244.

Nirenburg, S. (ed.). 1989. "Knowledge-based MT", *Special issue Machine Translation* vol.4, no 1 and 2, Kluwer Publishers, Dordrecht.

Nunberg, G & A. Zaenen. 1992. "Systematic Polysemy in Lexicology and Lexicography". *Proceedings of EURALEX'92*, University of Tampere.

Ostler, N. and S. Atkins. 1991. "Predictable Meaning Shift: some linguistic properties of lexical implication rules". Ed. Pustejovsky J. and Bergler S. *Lexical Semantics and Knowledge Representation*, Association for Computational Linguistics.

Peters, W., I. Peters, and P. Vossen. 1998a. "The Reduction of Semantic Ambiguity in Linguistic Resources". In: A. Rubio, N. Gallardo, R. Catro and A. Tejada (ed) *Proceedings of First International Conference on Language Resources and Evaluation*, Granada, 28-30 May 1998. 409-416

Peters, W., P. Vossen, P. Diez-Orzas, G. Adriaens. 1988b. "Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index. In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds), *Special Issue on EuroWordNet. Computers and the Humanities*, Volume 32, Nos. 2-3 1998. 221-251.

Pustejovsky, J. 1995. *The Generative Lexicon*, MIT Press, Cambridge MA.

Rodriquez, H., S. Climent, P. Vossen, L. Bloksma; A. Roventini, F. Bertagna, A. Alonge, W. Peters. 1998. "The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology". In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds), *Special Issue on EuroWordNet. Computers and the Humanities*, Volume 32, Nos. 2-3 1998. 117-152.

Vossen, P., L. Bloksma, C. Peters, A. Alonge, A. Roventini, E. Marinai, I. Castellon, T. Marti, G. Rigau, 1998. "Compatibility in Interpretation of Relations in EuroWordNet". In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds), *Special Issue on EuroWordNet. Computers and the Humanities*, Volume 32, Nos. 2-3 1998. 153-184.

Vossen, P., P. Diez-Orzas, W. Peters. 1997. "The Multilingual Design of EuroWordNet". In: P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) *Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, July 12th, 1997. 1-8.

Whitelock, P. 1991. "Shake-and-bake translation", Proceedings of Coling 1992.

[1] Further information on EuroWordNet and the builders can be found at http: //www.hum.uva.nl/~ewn.

[2] Parts of this section have also been published in Peters et al. 1998: 221-251.

[3] The actual internal organization of the synsets by means of semantic relations can still be recovered from the WordNet database, which is linked to the index as any of the other wordnets. In fact it is possible to get any semantic organization of a set of synsets in the index according to the hierarchy of any wordnet or ontology (in any language) linked to the index.

[4] In the real situation there are many more matches and translations. Here, we have listed just two of them to illustrate the example.