Piek Vossen, University of Amsterdam

Right or Wrong: Combining Lexical Resources in the EuroWordNet Project

Abstract

In the EuroWordNet-project we will construct wordnets from existing resources for Spanish, Dutch and Italian, similar to the Princeton WordNet1.5. Each of these wordnets will be developed as a separate language-internal system but each meaning will be linked to the closest meaning in WordNet1.5. In this way we hope to combine information from independently created resources, making the ultimate database more consistent and reliable, while keeping the richness and diversity of the vocabularies of the different languages. Two problems are discussed here: how to achieve overlap in the concepts that are represented in each wordnet and to interpret difference in the lexical semantic relations or configurations across the wordnets.

1. Introduction

WordNet is a database developed at Princeton University (Miller et al. 1990) with semantic relations between English words organized around the notion of *synsets*. Each synset comprises one or more word senses which are considered to be identical in meaning, together with a gloss which defines that meaning, e.g.: file2, data file1 = a set of related records kept together. This means that *file* in sense 2 is identical in meaning to *data file* in sense 1 and that the meaning is *a set of related records kept together*. Synsets can be related to each other by zero or more predefined relations, such as hyponymy, meronymy, cause, entailment (e.g. *binary file* is a kind of *file*, *record* is a part of *file*, etc.). The aim of the EuroWordNet project (LE2-4003) is to build a multilingual database with similar wordnets for several European languages (Dutch, Italian, and Spanish). These wordnets will be stored in a central lexical database system and the synsets will be linked to the closest synset in the Princeton WordNet1.5. Furthermore, we will merge the major concepts and words in the individual wordnets to form a common language-independent ontology (an ontology is the set of semantic relations between concepts). This will guarantee compatibility and maximize the control over the data across the different wordnets while language-dependent differences can be maintained in the individual wordnets. As builders will act the University of Amsterdam (Coordinator of the

project), the University of Sheffield, the Instituto di Linguistica Computazionale del CNR (Pisa), and the Fundación Universidad-Empresa (a co-operation of Universities of Barcelona and Madrid). The database will be tested in an information retrieval engine of Novell Linguistic Development in Antwerp. Further information on the project can be obtained from http: //www.let.uva.nl/CCL/EuroWordNet.html.

There are two extreme approaches to build such a multilingual database. Either the current Princeton WordNet is expanded with equivalence links from each synset to synsets in the other languages (the expand-model), or the wordnets are built separately in each language and are then linked to the most equivalent synset in WordNet1.5 (the merge-model). From a technical point of view, the expand-model seems less complex, guaranteeing the highest degree of compatibility across the different wordnets. The problem for the expand-model is however that the multilingual system will be highly biased by the Princeton WordNet. It will not only contain all the mistakes and gaps that are present in WordNet1.5 (just like any other dictionary) but it will also be structured by the (American)-English lexicalization of Western concepts. For these reasons, the EuroWordNet follows the technically-more-complex merge-model which starts from existing independently-developed resources (dictionaries and databases). This has the following advantages:

- Using existing resources is more cost-effective since a lot of work is already done.
- The different resources reflect the relations between words as separate language-internal systems. By combining existing resources it will be possible to maintain the language-dependent differences which will be erased when expanding from WordNet1.5.
- Experiments have shown that there is considerable variation in the way semantic information for equivalent words is coded within dictionaries and across dictionaries but that by combining resources a much higher degree of consistency can be achieved.

There are however two crucial factors in this approach which will be discussed in this paper:

- how to assure sufficient overlap in the coverage of the different wordnets and still maintaining language-specific properties of the relations (section 2)
- how to interpret differences found across the different wordnets (section 3)

The discussion will focus on nouns and examples will be taken from English and Dutch.

2. Ensuring overlap in coverage and language-sensitiveness of the wordnets

The wordnets will be developed separately by each builder for their own language from existing resources. In order to derive the explicit relations from the available machine readable dictionaries (MRDs) the following steps will be taken, using technology developed in the Acquilex-projects (BRA-303, BRA-7315) and Sift-project (LRE-62030):

- selection of a subset from the existing resources for which the relations will be specified.
- complete extraction of definition words needed for linking (making use of the available pattern matchers and definition parsers).
- determine the senses of the extracted definition words.
- interpret the syntactic relations with the definition words as hyponymy, meronymy or synonymy relations. This will result in a reorganization of the senses into synsets.
- link the synsets in each wordnet to the closest synsets in WN1.5.

After completion of these steps for the first subset the resulting wordnets will be loaded into the EuroWordNet Database in which they can be viewed and compared to check consistency across the resources. Using the Novell Wordnet Database (Diez-Orzas et al 1995), the major relations and hierarchy tops will be restructured and major mistakes that emerge from inspecting the wordnets will be corrected. The restructured tops and major (most frequently related) nodes of each individual wordnet will be merged to form a common top-ontology. The data will be restricted to nouns and verbs in English, Dutch, Italian and Spanish. We aim at a total set of 50,000 senses, correlating with about 20,000 most frequent words in the languages. The selection will have the following characteristics:

- there should be maximal overlap of the covered concepts across the different wordnets.
- the covered subset has to be generic: all frequent words of the language with their most frequent and common senses should be present.

- every parent concept that is needed to define a more specific concept should be present so that the introduction of new items does not require the addition of top-concepts.
- the subset should reflect language-specific lexicalization patterns.

The actual selection will take place in two phases. The first subset will be based on the defining vocabulary of each dictionary or resource from which the wordnets will be derived. This has the advantage that all words needed to link other words in the lexicon will be present in the selection of the wordnets, which will avoid technical complications. Since the defining vocabulary is probably not fully covered by the used dictionaries either missing defining words have to be added at the beginning. Furthermore, the words at the more general levels of the hierarchy are expected to be more difficult to define from a linguistic perspective. These words often have many vaguely distinguished meanings with a rather special linguistic usage. By linking all the words in the top most of the problematic cases will be handled. Any extension of the vocabulary in the wordnets will then involve the linking of more specific words to well-defined and delineated concepts in the wordnets; in other words we do not expect that extensions will introduce new hierarchical tops. Words belonging to the outer-shell of the language are also expected to be less linguistically-complex (although they may have a technical meaning).

Since the total set of defining words is rather large a more-specific selection will be made within the super-set of defining words. First, the most frequent defining words are selected and a top-ontology is created for these words (Bottom-up Selection I). The basic senses have to be selected and, if necessary, senses may have to be added to reflect the role of these words in defining the other words. Next, the children of the most frequent defining words will be included (Top-Down Selection II). By children we mean those words that have such a top word in their definition and which can be linked to it via one of the predefined relations. Finally, other defining words are added which have not yet been covered (Bottom-Up Selection III), where additional criteria may be used such as: presence of words in the bilingual dictionary; presence in available lists of basic words in the languages.

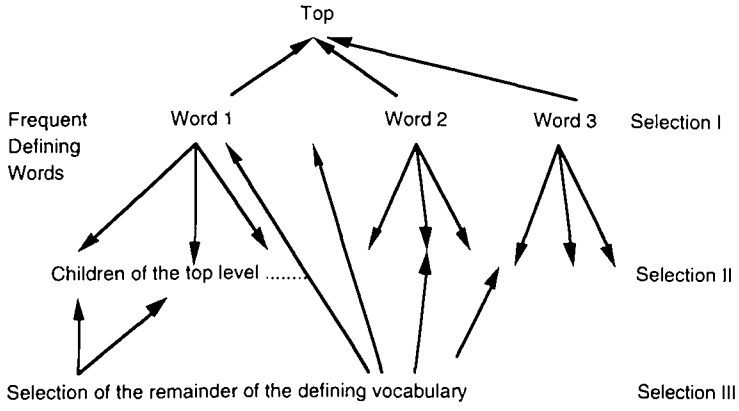

Figure 1: The selection of the first subset of the vocabulary

In the second building phase the subset will be extended along the following lines. After linking the first subset each site will have a list of WordNet synsets to which their entries have been linked. These lists will be exchanged and compared. Those WordNet synsets that are not present in a site's list but are present in the lists of the other sites will be used to generate the first extension (via the bilingual dictionaries). To achieve sufficient overlap and compatibility we will also make use of the consistency-checking and wordnet-comparison mechanisms that will be implemented in the EuroWordNet database. Extreme differences across wordnets (e.g. in lexical density) or incompatibility of redundant relations will be inspected (see the next section).

To achieve sufficient genericity of the wordnets frequency lists will be extracted from a diverse range of corpora. Other criteria will be the word length, morphological complexity and the degree of polysemy. However, we expect that the defining vocabulary will mostly coincide with the more frequent words in daily language use. The extension from the corpora is probably minimal. More difficult is it to determine the frequency of the senses of the words that are included. In principle, we will exclude obscure and rare senses on the basis of labelling in the dictionaries from which the wordnets are derived and by inspection. These senses may be added later on when they can be linked to other senses as specific variants (coded for register, dialect, as grammatical variants, etc.) of present concepts.

Whereas the two previous strategies will lead to the construction of the hierarchy in a bottom-up fashion (selected words are linked to more general levels), for the third extension the hierarchy will be traversed top-down so that 'missing siblings' of nodes in the hierarchy can be

added (e.g. *cat* is linked to *animal* but *pet* is not included in the subset). Using this method more complete lexicalisation patterns of concepts in a particular language will be covered (which is not guaranteed by the above strategies). These lexicalisations include language-specific phenomena and different types of variants (possibly also the less frequent and basic senses of the frequent words that have been omitted at the beginning). In addition to words expressing a concept we will investigate the possibility to include multi words, typical phrases and expressions linked to concepts.

3. Differences in relations across wordnets

Once such a multilingual database is derived the relations between the wordnets can be compared using bilingual dictionaries. The following situations can occur:

1) there is a mismatch according to a bilingual dictionary between two word senses in a language pair (Soler and Marti 1993, Vossen 1993, Copestake et al 1995).
2) there is a mismatch in the wordnet relations of two senses which have some equivalence relations according to a bilingual source (Vossen 1991, Ide and Veronis 1994, Vossen 1995).
3) a combination of 1) and 2).

3.1 Different types of equivalence relations

The mismatches listed in 1) can be categorized on the basis of the kind of information given in the bilingual dictionary. In the case of noun senses the following types can be distinguished (examples are taken from the van Dale bilingual dictionaries: Dutch-English (Martin and Tops 1986, short reference VDE), and English-Dutch (Martin and Tops 1984, short reference VED):

- gap: there is no translation for a word sense according to a bilingual dictionary. For example, the Dutch word *lawaaisaus* ('a thin and watery sauce') is not included in the VDE and *Madeira sauce* which is the translation given for *Madeirasaus* in the VDE is not included in WordNet1.5.

- the translation is marked with a grammatical label and/or has an inflected form, e.g.:

Entry Word	Label	Translation VDE	Label
afrastering		railings	plural
proteine	plural	protein	singular
bijenvolk (lit. 'beespeople')		bees	

- the translation is not an entry in the other wordnet but a derivation of an entry, e.g.:

Entry Word	Translation VDE	EntryWord	Translation VED
stomheid	dumbness	wicklet	deurtje (deur+diminutive'tje')
gewrongenheid	forcedness	beer	biertje (bier+diminutive 'tje')

- the translation is a phrase: adjective-noun combinations (a); prepositional adjuncts (b); relative clauses (c); compound translations (d):

	Entry Word	Translation VDE		Entry Word	Translation VDE
a	landbouwschap	agricultural board	c	bluswater	(fire extinguishing) water
	aardbeienneus	red, bulbous nose		aardappelpoter	potato planting machine
b	andragoloog	specialist in adult education		cultuur-pessimist	pessimist who sees little future in culture
	antipodespel	juggling with the feet			
	dropje	piece of liquorice	d	dierenkliniek	animal clinic
	hout	piece of wood		dierentaal	animal language
	woordgroep	group of words			

- there are multiple translations for a word sense:

Entry Word	Translation VED
leg	been (of a human); poot (of an animal)

- the translation is labelled with a register or dialect label

In all these cases the bilingual dictionary may express a lexicalized difference between the two languages but not all difference are equally relevant for the information coded in the wordnets. Register and dialect labels will not effect the semantic relationship between words. In fact we can state that only those differences are relevant that somehow affect the semantic relations stored in the wordnets. In the case of nouns this will mostly be hyponymy, meronymy and synonymy. The structurally distinguished translations shown above can be classified semantically as follows:

i. hyponymy relation between the source entry and the target entry:

• adjective + noun:	aarbeienneus	-hyperonym--> nose
• noun + noun:	dierentaal	-hyperonym--> language
• noun + relative clause:	aardappelpoter	-hyperonym--> machine
• noun + PP:	andragoloog	-hyperonym--> specialist
	noordeling	-hyperonym--> potato
• some derivations	wicklet	-hyperonym-->deur<door>

ii. meronymy relation between the source entry and the target entry:

• noun + PP:	woordgroep	-holonym--> word
	meubel	-meronym-->furniture
• some derivations:	beer	-holonym--> bier
• some inflected forms:	bijenvolk	-holonym-->bee

It will be clear that not all structural properties can be directly interpreted as a specific semantic relation. Only particular prepositional-adjuncts can be interpreted as expressing meronymy relations (only those having relational heads such as *part*, *piece*, *group*, *member* etc.), and something similar holds for the derivations. These interpretations can however be specified in much the same way as explicit relations can be extracted from monolingual dictionary definitions. In the case of inflected forms it is necessary to know whether the plural form is functional or not. In the case of English-Dutch pairs such as 'oats/havermout', 'bran/zemelen', where the underlined examples are pluralia tantum and the equivalents are singularia tantum, the plurality does not correspond with multiplicity in denotation. Such (un)functionality of plural form (Vossen 1995) may be determined by the correlation between the plural/singular form and the structure of the monolingual definitions.

Finally, there may be a difference in the status of the hyponymy relation expressed by phrasal translations (or the absence of a translation in the case of a translation gap). Whereas in some cases a word in one language names a more specific type of denotation, in other cases a lexicalization expresses a conceptualization. In all the following examples a concept is lexicalized in Dutch but not in English:

	Entry Word	**Phrasal Translation**
a	theewater	water used for making tea
	koffiewater	water used for making coffee
	bluswater	water used to extinguish fire
b	citroenjenever	lemon geneva
	rood	red-currant gin

The examples in a) do not refer to different types of *water* but refer to water used for a particular purpose: limiting the perspective or expressing a specific conceptualization. In the example at b), on the other hand, specific types of *gin* are mentioned which are not lexicalized in English. This difference in hyponymy relation is important to measure the overlap in concepts across the wordnets. Whereas lexicalized conceptualizations can be seen as language-specific information of a wordnet, differences in denotational coverage should be avoided. This difference is however difficult to automatically extract from the structural properties of the translations in the bilingual dictionaries.

3.2 Differences in lexical configurations of wordnets

In the case of a simple translation (no labels, inflection or phrase) we can assume that there is also a synonymy relation between two word senses in two wordnets. In principle we can say that in that case all relations should be parallel, as shown in Figure 2.

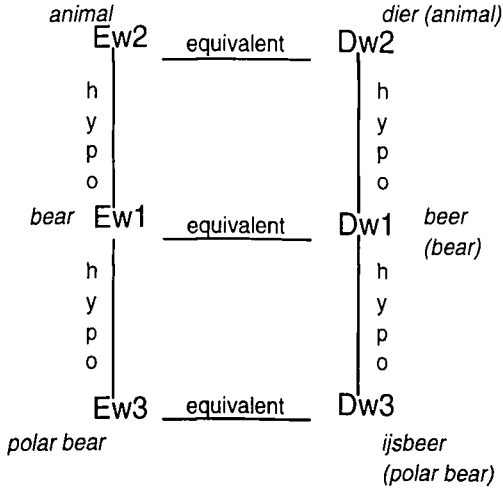

Figure 2: Parallelism in wordnet configurations

If an English word Ew1 (*bear*) has a simple equivalence relation with a Dutch word Dw1 (*beer*) then we expect that English words (e.g. Ew2 and Ew3) related to Ew1 in the English wordnet which are equivalent to Dutch words (e.g. Dw2 and Dw3) related to Dw1 in the Dutch wordnet have the same semantic hyponymy or meronymy relation. However, regardless of an equivalence relation between two words their relations

in the wordnets can still differ in various ways because the way of defining words in dictionaries may vary (Vossen 1995):

- a different selection of features: different properties have been chosen to capture the meaning of a concept either through an omission or simply because not all properties can be given within the limited space and time for writing definitions.
- a different priming of features: given a set of features that are essential for the meaning of a word, a lexicographer can make a different choice in selecting one feature as the head and the other features as the differentiae.
- a different abstraction of features: given a set of features a lexicographer can always choose to classify it at a more abstract level and to specify the extra discriminating features as differentiae.

These differences are not errors (in the sense that a *cat* is defined as a *vehicle*) but are either due to the fact that only one classification scheme or perspective can be expressed in a definition where several may apply, or that there are multiple ways to express the same classification scheme. In this respect, recursively adding linked wordnets may lead to an improvement of the relations. In Figure 3, for example, part of the lexical semantic configuration for *dog* is represented as derived from the Longman Dictionary of Contemporary English (Procter 1978, LDOCE), the Van Dale monolingual Dutch dictionary (van Sterkenburg and Pijnenburg 1984) and WordNet1.5. We see here that in LDOCE *pet*, *mammal* and *dog* are all directly defined as subtypes of *animal* (probably due to the use of the controlled vocabulary). The relation between *dog* on the one hand and *mammal* and *pet* on the other hand is not indicated. In Van Dale and WordNet, we see that each expresses one of these more specific relations but none of them expresses both (because only one perspective or conceptualization is given in a definition and in WordNet1.5).

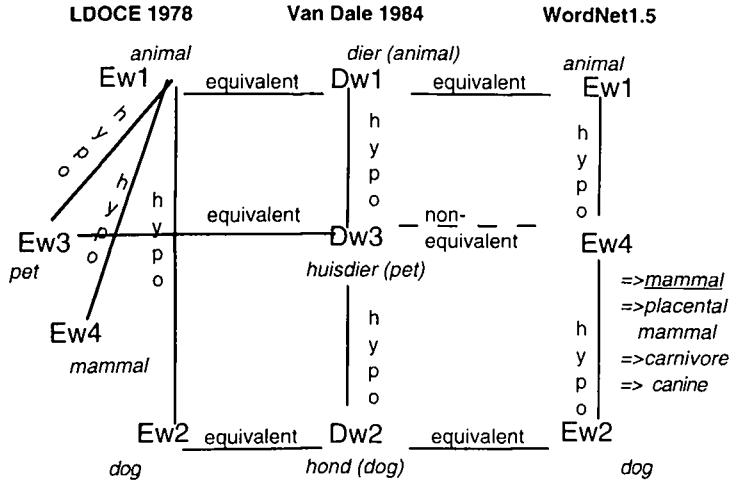

Figure 3: Variation in lexical configurations across wordnets

Either the equivalence relations derived from the bilingual dictionaries are incorrect, or the configurations are wrong or incomplete. Bilingual dictionaries may be inadequate in that a more specific word is translated with a more general word when there is no direct equivalence. Especially in the case of denotational gaps as we have seen above (e.g. *rood* ('red-currant gin')), a bilingual-dictionary may give a more general term which would be more appropriate than the full phrase as a translation. In this example however, the wordnet configurations are incomplete. Assuming that the equivalence relations are correct, we can infer from the equivalence relations between Dw3 and Ew3 that *pet* must be more general than *dog* unless the hyponymy relation between Dw3 and Dw2 in Van Dale is wrong (it could have been synonymy or co-hyponymy). Something similar can be said for *mammal* in WordNet1.5 and LDOCE. Either *mammal* is skipped as a level for *dog* or it should be synonymous with *dog* or a co-hyponym. From an incidental comparison between two wordnets it may be difficult to infer what relation is right or wrong and not all cases will lead to such an obvious conflict. However, when more and more wordnets are added, repeated parallelism can be strengthened and isolated relations (only occurring in a single wordnet) will be more suspicious. If another wordnet confirms the relations expressed in Van Dale and WordNet1.5 this may lead to a restructuring of the relations so that *dog* is related to both *mammal* and *pet* in all wordnets.

Because dictionary definitions only take a single perspective as a starting point we see that components are sometimes defined with a

hyponymy relation and sometimes with a meronymy relation. Especially in the case of meronymy relations, this means that derived configurations will be very incomplete. In Figure 4 we see that both LDOCE and Van Dale define *arm* and *leg* ('been' in Dutch) as subtypes of *limb* and Dutch *ledemaat*, whereas *head* and *hoofd* are directly defined as meronyms of *body* and *lichaam* respectively.

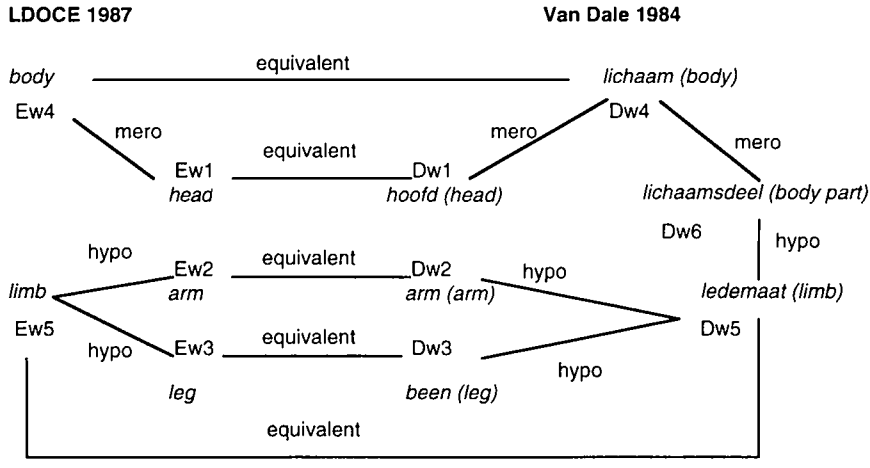

Figure 4: Hyponymy and meronymy relations across wordnets

Unfortunately, *limb* is in LDOCE defined as a *leg*, *arm*, or *wing* and is not related to *body*. However, since in Van Dale *ledemaat* ('limb') is defined as a hyponym of *lichaamsdeel* ('body part') and likewise indirectly as a meronym of *lichaam* we can recreate the missing link on the basis of the equivalence relations.

Note that the previous example also illustrates another interesting phenomena, namely that *arm*, *leg* and *head* in English can name body parts of both humans and animals, whereas in Dutch the above examples are only used for *humans* (and *horses*). For animals (except *horses*) *kop* ('head') and *poot* ('leg') should be used. Since the bilingual dictionary gives two translations for *head* (*hoofd* and *kop*) and *leg* (*been* and *leg*) this should lead to a multiple matching (*human and animal bodies*). This is not in conflict with the general holonym *body* in English and it therefore should not be a problem. More strongly, we can infer from the comparison that the English body parts generalize over humans and animals. If the translation relation between words of the two languages are differentiated in terms of the same relations that are distinguished in the wordnets we can thus use this information to check the lexical

semantic configurations in each wordnet and derive more specific information.

4. Conclusions

We have described a strategy to independently build language-internal systems of semantic relations or wordnets for the generic, general part of several languages, while ensuring overlap and compatibility and maximizing systematicity and control across these wordnets. Furthermore we saw that differences between the wordnets can either be reflected by differences coded in the bilingual dictionaries (a complex translation-relation) or by differences in the lexical configurations. When there is a parallelism in the lexical configurations across two wordnets we expect that the translation relations should be straight-forward as well. When there is a divergence of the lexical configurations of two wordnets the difference should be reflected by the translation-relation in the bilingual dictionary as well. If a divergence of lexical configurations is confirmed by a complex translation-relation coded in the bilingual dictionaries it strongly suggests that the difference is language-specific. If these two correlations do not hold either the lexical configurations are incomplete or the information in the bilingual dictionaries. Configurations reflected in a large number of wordnets will increase their credibility. Constraints in the equivalence relations and the corresponding lexical configurations will be stored as consistency-checking constraints in the EuroWordNet database, so that by combining the individually-built wordnets the redundancy can be exploited to automatically look for suspicious configurations and to deduce a measure of consistency across the wordnets.

References

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", in: Dorr and Klavans (eds.), *Special Issue on Building Lexicons for Machine Translation, Machine Translation*, 9-3, Kluwer, 33–69.

Díez-Orzas P., et al, 1995. The Novell ConceptNet, Internal Report, Novell Belgium N.V.

Ide, N. and J. Veronis 1994 "MachineReadable Dictionaries: What have we learned, Where do we go?", in: N. Calzolari and C. Guo (eds.), *Proceedings of the post-Coling94 international workshop on directions of lexical research, August15–17, Beijing*, 137–146.

Martin, W. and G.A.J. Tops (eds.) 1984. *Groot Woordenboek Engels-Nederlands*, Van Dale Lexicografie, Utrecht.

Martin, W. and G.A.J. Tops (eds.) 1986. *Groot Woordenboek Nederlands-Engels*, Van Dale Lexicografie, Utrecht.

Miller G.A., Fellbaum, C., Gross, D. & K.J. Miller 1990. "Introduction to WordNet: An On-line Lexical Database", in: *International Journal of Lexicography*, Vol 3, No.4 (winter 1990), 235–244.

Proctor, P. (ed.) 1978. *The Longman dictionary of contemporary English.* London: Longman.

Sterkenburg J. van, and W.J.J. Pijnenburg (eds.) 1984. *Groot woordenboek van hedendaags Nederlands*, Van Dale Lexicografie, Utrecht.

Soler, C. and M. A. Marti 1993 Dealing with Lexical Mismatches - Esprit BRA-7315. Acquilex2 Working Paper. 4.

Vossen P. 1991. Comparing noun-taxonomies cross-linguistically, Esprit BRA-3030 Acquilex Working Paper 014. University of Amsterdam.

Vossen P. 1993. Extracting equivalence relations for a multilingual lexical knowledge base, Esprit BRA-7315 Acquilex2 Working Paper 014, University of Amsterdam.

Vossen P. 1995. *Grammatical and conceptual individuation in the lexicon*, PhD Thesis, University of Amsterdam, Ifott 15.