

# **EuroWordNet Subset2 for Dutch, Spanish and Italian**

**Version 4, Final**

November 27, 1998

Contributors:

Piek Vossen, Laura Bloksma, University of Amsterdam

Salvador Climent, Maria Antonia Marti, Mariona Taule, Universitat de Barcelona

Julio Gonzalo, Irina Chugur, M. Felisa Verdejo, UNED

Gerard Escudero, German Rigau, Horacio Rodriguez, Universitat Politecnica de Catalunya

Antonietta Alonge, Francesca Bertagna, Rita Marinelli, Adriana Roventini, Luca Tarasi, Istituto di Linguistica del CNR, Pisa



**Deliverable D027, D028, WP3, WP4  
EuroWordNet, LE2-4003**

Identification number	LE-4003-D027-D028
Type	Document and Lingware
Title	EuroWordNet Subset2 for Dutch, Spanish and Italian
Status	Final
Deliverable	D-027, D-028
Work Package	WP3 and WP4
Task	T4
Period covered	April 1998 – October 1998
Date	November 27, 1998
Version	4
Number of pages	77
Authors	<ul style="list-style-type: none"> <li>⇒ Piek Vossen, Laura Bloksma, University of Amsterdam</li> <li>⇒ Salvador Climent, Maria Antonia Marti, Mariona Taule, Universitat de Barcelona</li> <li>⇒ Julio Gonzalo, Irina Chugur, M. Felisa Verdejo, UNED</li> <li>⇒ Gerard Escudero, German Rigau, Horacio Rodriguez, Universitat Politecnica de Catalunya</li> <li>⇒ Antonietta Alonge, Francesca Bertagna, Rita Marinelli, Adriana Roventini, Luca Tarasi, Istituto di Linguistica del CNR, Pisa</li> </ul>
WP/Task responsible	PSA/FUE
Project contact point	<p>Piek Vossen  University of Amsterdam  Spuistraat 134  1012 VB Amsterdam  The Netherlands  tel. +31 20 525 4669  fax. +31 20 525 4429  e-mail: <a href="mailto:Piek.Vossen@hum.uva.nl">Piek.Vossen@hum.uva.nl</a></p>
EC project officer	Ray Hudson
Status	Public
Actual distribution	Project Consortium, the EuroWordNet User Group, the world via <a href="http://www.hum.uva.nl/~ewn">http://www.hum.uva.nl/~ewn</a> .

Supplementary notes	n.a.
Key words	Linguistic Resources, Multilingual Wordnets, Language Engineering
Abstract	This deliverable describes the wordnets for Dutch, Italian and Spanish after the second building round. The first versions of the wordnets have been compared and the results of the comparison have been used to improve the wordnets. The new comparison, described in this document, indicates a major progress towards the final state of the wordnets.
Status of the abstract	Final
Received on	
Recipient's catalogue number	

## Executive Summary

This deliverable describes the results of extending the core wordnets for Dutch, Italian and Spanish, described in D014D015 (Vossen et al. 1998), to the final set of synsets and relations, and comparing these complete sets. The complete set of synsets is called Subset2 and it should consist of 50K word senses per language, which more or less corresponds with 25-35K synsets. The extension to Subset2 has been based on the comparison of the core wordnets (Subset1). From this comparison, we derived the following priorities for improvement:

- improve the balancing of 1stOrderClusters (Dutch and Italian)
- extend with missing top-frequent Parole entries (Dutch, Italian, Spanish)
- extend the coverage (Dutch)
- check translations of extremely long hyponymy chains (especially Dutch verbs)
- check sequences with 1 gap (Italian, Spanish and Dutch)
- extend translations (Italian)
- improve translation heuristics (Spanish and Dutch)

In this document we describe the individual strategies to reach these goals and we give for each language the tables that summarize the current wordnets (section 2). Furthermore, we describe the results of comparing them in terms of the distribution or clustering over the top-ontology of 63 semantic distinctions (section 3) and by a graph-comparison of the hyperonym-translation chains across the wordnets (section 4).

We concluded that the wordnets have progressed considerably compared to Subet1 and are reaching their final state. Quantitatively, we made the following progress:

- all wordnets have reached their final quantitative state: 23-35K synsets and 45-55K word meanings, which is about 1/3 of the size of WordNet1.5.
- all 3 wordnets include the most frequent entries from the corresponding Parole lexicons in the languages.
- the average number of relations (average of 2.4) has become more balanced as compared to Subset1, and is slightly higher than WordNet1.5 (1.4).
- clustering across the top-ontology is balanced across WordNet1.5, Spanish and Dutch.

Qualitative improvements with respect to Subset1 are:

- the number of synonyms or synset variants per synset has become more balanced across the wordnets.
- the ILI-intersection by all 3 languages has increased 7 times, and language-pairs have intersections that make up 20 to 40% of the total set of concepts represented.
- the length of the ILI chains got more balanced.
- the overlap of partial chains increased considerably (up to 4 times).
- the overlap of partial chains with one gap increased even more (up to 4 times for nouns and up to 60 times for verbs)

We also investigated the effect of Composite ILIs on the compatibility of the ILI-chains. We concluded that the Composite ILIs:

- raised the complexity of the chains
- increased the overlap of partial chains: up to 10% for nouns and up to 24% for verbs
- increased the overlap of partial chains with one gap: up to 7% for nouns and up to 13% for verbs

We also concluded that more work needs to be done with respect to:

- the quality of equivalence relations (this would eliminate extremely long ILI-chains in e.g. the Dutch wordnet).
- the number of equivalence relations in the Italian wordnet.

Finally, we stated that complete overlap is both impossible and not desirable. The size of the wordnets in EuroWordNet is 1/3 of the size of WordNet1.5, which simply makes it impossible to have full overlap of hierarchies. However, in many cases there are also structural differences between the wordnets at high levels in the hierarchy as a matter of choice. This also seriously affects the compatibility measure of the longer chains.

## Table of Contents

1. General approach for building the wordnets .....	9
2. Overview of Subset2 for Dutch, Spanish and Italian.....	11
2.1 Subset2 for the Dutch wordnet.....	11
2.1.1. Increase of coverage.....	11
2.1.2. Improving the equivalence relations .....	14
2.1.3. Overview tables for the Dutch wordnet (Subset2) .....	21
2.2. Subset2 for the Italian wordnet .....	25
2.2.1. Increase of coverage.....	25
2.2.2. Building synsets, restructuring taxonomical chains and encoding additional relations .....	26
2.2.3. Encoding equivalence relations .....	28
2.2.3. Overview tables for the Italian wordnet (Subsets1+2) .....	30
2.3. Subset2 for the Spanish wordnet .....	33
2.3.1. Improving the quality of Subset 2: work on the content of synsets.....	34
2.3.2. Improving the quality of Subset 2: creation of new relations between synsets.....	37
2.3.3. Overview tables for the Spanish wordnet (Subset 2) .....	39
2.4. Quantitative conclusions.....	43
3. Coverage of Subet2 over top concept clusters .....	46
4. Comparison of Subset2.....	51
4.1. Introduction.....	51
4.2. Evaluation of individual wordnets.....	52
4.3. Global evaluation .....	55
4.4. Comparison with composite ILIs.....	60
5. Conclusions .....	70
References .....	71
Appendix I: Explanations of Gaps in the Dutch wordnet compared to Spanish, Italian and English.....	72
Appendix II Projection of complete chains on the Dutch, Italian and Spanish wordnets.....	74
Appendix III Projection of partial chains on the Dutch, Italian and Spanish wordnets .....	75

## List of Tables

Table 1: Construction of the Dutch nominal wordnet .....	13
Table 2: Synsets without translation in the Dutch database.....	14
Table 3: Directly matching entries between the Dutch database and WordNet1.5. ....	14
Table 4: New nominal translations generated by direct matches with WordNet1.5.....	14
Table 5: Translations by Reversed English-Dutch dictionary and Replacing Spaces and Hyphens.....	15
Table 6: Synsets without translation in the Dutch wordnet .....	15
Table 7: Synsets classified with top-concepts.....	18
Table 8: Conversion of WordNet1.5 Lexicographer's file codes to EuroWordNet top-concepts.....	19
Table 9a: Automatic matching results for nouns .....	20
Table 9b: Automatic matching results for verbs .....	20
Table 10: Subset2 Overview NL.....	21
Table 11: Language Internal Relations NL.....	22
Table 12: Status of the Language Internal Relations NL.....	23
Table 13: Equivalence Relations NL.....	23
Table 14: Quality of the equivalence relations .....	24
Table 15: Synsets not mapped to WN1.5 .....	28
Table 16: Synsets mapped using different methodologies.....	29
Table 17: Italian Subsets1+2 Overview.....	30
Table 18: Italian Language Internal Relations .....	31
Table 19: Italian Equivalence Relations* .....	32
Table 20: Spanish Relations derived from Suffixes .....	38
Table 21 : Subset2 Overview ES .....	40
Table 22: Language Internal Relations ES .....	41
Table 23: Equivalence Relations ES .....	42
Table 24: Reliability of Equivalence Relations ES .....	42
Table 25: Subset2 Overview: NL, ES, IT.....	43
Table 26: Overview of senses and synsets for WordNet1.5. ....	43
Table 27: Overview of Language Internal Relations.....	44
Table 28: Distribution of different relations in WordNet1.5 .....	45
Table 29: Overview of Equivalence Relations.....	45
Table 30: Nominal Synsets clustered as 1stOrder Concepts.....	47
Table 31: Nominal Synsets clustered as 2ndOrder Concepts.....	48
Table 32: Verbal Synsets clustered as 2ndOrder Concepts .....	49
Table 33: Nominal Synsets clustered as 3rdOrder Concepts .....	49
Table 34: Subset2 clustered over the WordNet1.5 Lexicographer's file codes .....	50
Table 35 ILI chains for nouns.....	52
Table 36 ILI chains for verbs.....	52
Table 37 ILI chains (total).....	52
Table 38 Frequencies and ratios of noun chains / length /language .....	53
Table 39 Frequencies and ratios of verb chains / length /language.....	54
Table 40 Coverage of noun ILI records.....	55
Table 41 Coverage of verb ILI records .....	55
Table 42 Coverage of ILI records (total).....	56
Table 43 Coverage of complete noun chains projected over WN1.5 structure .....	56
Table 44 Coverage of complete verb chains projected over WN1.5 structure.....	56
Table 45 Coverage of partial noun chains of NODES projected over WN1.5 structure.....	57
Table 46 Coverage of partial noun chains of EDGES projected over WN1.5 structure .....	57
Table 47 Coverage of partial VERB chains of NODES projected over WN1.5 structure .....	58
Table 48 Coverage of partial VERB chains of EDGES projected over WN1.5 structure .....	58
Table 49 Comparison of partial coverage of WN1.5 chains by the intersection of WNs between subset1 and subset2..	58
Table 50 Coverage of partial noun chains of NODES with 1 gap projected over WN1.5 structure.....	59
Table 51 Coverage of partial NOUN chains of EDGES with 1 gap projected over WN1.5 structure.....	59
Table 52 Coverage of partial VERB chains of NODES with 1 gap projected over WN1.5 structure .....	59
Table 53 Coverage of partial VERB chains of EDGES with 1 gap projected over WN1.5 structure.....	59
Table 54 ILI chains for nouns (composite ILIs).....	60
Table 55 ILI chains for verbs (composite ILIs) .....	60
Table 56 Frequencies and ratios of nominal chains / length /language (composite ILI).....	61
Table 57 Frequencies and ratios of verbal chains / length /language (composite ILI) .....	62

Table 58 Coverage of complete noun chains projected over WN1.5 structure (composite ILI) .....	65
Table 59 Coverage of complete verb chains projected over WN1.5 structure (composite ILI) .....	66
Table 60 Coverage of partial noun chains of NODES projected over WN1.5 structure (composite ILIs) .....	67
Table 61 Coverage of partial noun chains of EDGES projected over WN1.5 structure (composite ILIs).....	67
Table 62 Coverage of partial verb chains of NODES projected over WN1.5 structure (composite ILIs) .....	68
Table 63 Coverage of partial verb chains of EDGES projected over WN1.5 structure (composite ILIs).....	68
Table 64 Coverage of partial noun chains of NODES with 1 gap projected over WN1.5 structure (composite ILIs) ....	69
Table 65 Coverage of partial noun chains of EDGES with 1 gap projected over WN1.5 structure (composite ILIs).....	69
Table 66 Coverage of partial verb chains of NODES with 1 gap projected over WN1.5 structure (composite ILIs) .....	70
Table 67 Coverage of partial verb chains of EDGES with 1 gap projected over WN1.5 structure (composite ILIs).....	70
Table 68 Coverage of complete noun chains projected over Dutch wordnet structure .....	74
Table 69 Coverage of complete verb chains projected over Dutch wordnet structure .....	74
Table 70 Coverage of complete noun chains projected over Italian wordnet structure .....	74
Table 71 Coverage of complete verb chains projected over Italian wordnet structure.....	74
Table 72 Coverage of complete noun chains projected over Spanish wordnet structure.....	74
Table 73 Coverage of complete verb chains projected over Spanish wordnet structure.....	74
Table 74 Coverage of partial noun chains of NODES with 1 gap projected over Dutch wordnet structure .....	75
Table 75 Coverage of partial noun chains of EDGES with 1 gap projected over Dutch wordnet structure.....	75
Table 76 Coverage of partial verb chains of NODES with 1 gap projected over Dutch wordnet structure .....	75
Table 77 Coverage of partial verb chains of EDGES with 1 gap projected over Dutch wordnet structure .....	75
Table 78 Coverage of partial noun chains of NODES with 1 gap projected over Italian wordnet structure.....	76
Table 79 Coverage of partial noun chains of EDGES with 1 gap projected over Italian wordnet structure .....	76
Table 80 Coverage of partial VERB chains of NODES with 1 gap projected over Italian wordnet structure .....	76
Table 81 Coverage of partial VERB chains of EDGES with 1 gap projected over Italian wordnet structure.....	76
Table 82 Coverage of partial noun chains of NODES with 1 gap projected over Spanish wordnet structure .....	76
Table 83 Coverage of partial noun chains of EDGES with 1 gap projected over Spanish wordnet structure.....	77
Table 84 Coverage of partial VERB chains of NODES with 1 gap projected over Spanish wordnet structure.....	77
Table 85 Coverage of partial VERB chains of EDGES with 1 gap projected over Spanish wordnet structure .....	77

## 1. General approach for building the wordnets

The EuroWordNet database is built (as much as possible) from available existing resources and databases with semantic information developed in various projects. In general, the wordnets are built in two major cycles. Each cycle consists of a building phase and a comparison phase:

1. Building a wordnet fragment
  - 1.1. Specification of an initial vocabulary
  - 1.2. Encoding of the language-internal relations
  - 1.3. Encoding of the equivalence relations
2. Comparing the wordnet fragments
  - 2.1. Loading of the wordnets in the EuroWordNet database
  - 2.2. Comparing and restructuring the fragments
  - 2.3. Measuring the overlap across the fragments

The building of a fragment is done using local tools and databases which are tailored to the specific nature and possibilities of the available resources. The available resources differ considerably in quality and explicitness of the data. Whereas some sites have the availability of partially structured networks between word senses, others start from genus words extracted from definitions that still have to be disambiguated in meaning.

The first wordnet subsets have been created from a set of 1024 Base Concepts. These Base Concepts play an important role in at least two wordnets, where importance is measured in terms of numbers of relations and position in the hierarchy. The Base Concepts have been represented (as far as possible) by synsets in Dutch, Italian and Spanish, and have been extended with other synsets that are important in these languages. These sets have been encoded and extended to form the first subset (minimally 10,000 synsets). These first subsets represent the cores of the different wordnets on which the meanings of more specific concepts depend. D014D015 (Vossen et al. 1998) contains a detailed description of the first subset and the results of comparing them.

The comparison of the core wordnets showed a need for the following improvements:

- balance the distribution of 1stOrderClusters (Dutch and Italian)
- extend the wordnets with top-frequent Parole entries that are missing (Dutch, Italian, Spanish)
- extend the coverage to the full size (50K word senses and 30L synsets, especially Dutch)
- check the translations of extremely long hyponymy chains (especially Dutch verbs)
- check sequences with 1 gap (Italian, Spanish and Dutch)
- extend the translations (Italian)
- improve translation heuristics (Spanish and Dutch)

This deliverable describes the work of extending the core wordnets to the final set of synsets and relations, and comparing the result. The languages covered are Dutch, Spanish and Italian. The total set of synsets aimed at for Subset2 is 50K word meanings, which more or less corresponds with 25-30K<sup>1</sup> synsets (20K nouns and 10K verbs). For each of these synsets, the following information has to be minimally specified:

- Hyperonym
- Synonyms (synset members)
- Equivalence relations to the Inter-Lingual-Index (WordNet1.5)

Optionally, any other relation has been added. The addition of other relations first of all depends on the relevance of the relation for the synset. Secondly, we have been limited by the project resources. Given the project-funding, it is not possible to comprehensively encode all relevant relations.

The subsets for Dutch, Italian and Spanish are further described in 3 sections:

- section 2: for each language, a description of the methodology followed, tables with the number of entries, senses, and synsets covered and the number and kind of relations encoded for each language.
- section 3: distribution of the vocabulary over the different top-ontology clusters.
- section 4: comparison of Subset2 using the FUE graph-comparison software.

---

<sup>1</sup> The number of correlating synsets depends on the definition of synonymy that is used within a wordnet. Since these synonyms may be extracted automatically the ratios may differ across the wordnets.

## 2. Overview of Subset2 for Dutch, Spanish and Italian

### 2.1 Subset2 for the Dutch wordnet

In the second building phase for the Dutch wordnet, we focussed on:

1. extending the core wordnet to the full size of 50K word senses (approximately 30,000 synsets)
2. improving the overlap across the wordnets
3. improving the quality of the equivalence relations

Task 1 and 2 are closely related and will be discussed in the subsection 2.1.1. The work on the equivalence relations is described in section 2.1.2. Section 2.1.3. gives the overview tables for the Dutch wordnet.

#### 2.1.1. Increase of coverage

The core wordnet for Dutch (the first subset) covered 9,588 synsets: 5,917 nominal synsets, 3,282 verbal synsets and 389 adjectival/adverbial synsets. This core wordnet had to be extended to approximately 30,000 synsets (20,000 nominal synsets and 10,000 verbal synsets).

For the verbs this implied extending the set to the complete lexicon of 9,125 synsets made available by Van Dale. These synsets contain 14,278 senses and 8,868 entries. The verbal wordnet has 100% overlap with the Parole lexicon for Dutch for entries with a frequency above 100.<sup>2</sup> The verbal wordnet for Dutch has two verbal tops "zijn" (to be) for static verbs and "gebeuren" (to happen) for dynamic verbs. All verbal synsets are connected to one of these tops via at least one hyperonym link. In addition, they may have other links (see tables in section 2.1.3.).

The coverage of the nominal wordnet has been increased by several measures:

1. Significantly low coverage of ontological clustering, as followed from the top-concept clustering. This mainly involved 1stOrderEntities.
2. Extending the vocabulary with missing Parole entries (all senses) with a frequency above 100.
3. Investigation of significant ILI chain gaps: these are translations of hierarchy nodes covered by the other wordnets but not in the Dutch wordnet.
4. Inclusion of all synsets with 10 or more relations
5. Inclusion of all synsets with 1 or 2 automatically derived translations

Coverage of the First Subset has been measured by clustering of synsets per Top Concept. This clustering is achieved by collecting the Top Concepts for all translations to WordNet1.5 synsets or hyperonyms of these translations in WordNet (see section 3).

---

<sup>2</sup> The Dutch Parole lexicon is developed by the Instituut voor Nederlandse Lexicografie (INL). The lexicon contains morpho-syntactic information for the most frequent words taken from corpora (Kruyt 1998). The INL has compared the Dutch core wordnet with their lexicon for different frequency clusters to measure the overlap.

The comparison of the 2ndOrderEntity Clusters did not show any unbalanced clusters across the sites. No balancing of the coverage was needed. We therefore worked on the 1stOrderEntity clusters for which the Dutch subset was significantly less filled in then the other sites: Animal; Creature; Function; Garment; Gas; Group; Human; Living; Occupation; Plant; Software. Since the size of the final wordnet is 1/3rd of the size of WordNet1.5, and ideal balancing would result in 1/3rd of the coverage of WordNet1.5. for each cluster. Except for Group, we have maximally balanced the clusters by adding more specific concepts. In the case of Animal and Plant, it is however impossible to get the same distribution because many exotic classes in WordNet1.5 are not available in our resource. In section 3, the results of the current balancing is discussed.

To get an overview of the conceptual coherence we exported the complete hierarchy as sorted flat chains:

```
iets 1 middel 2 vervoermiddel 1 voertuig 1 motorrijtuig 1 auto 1 automaat 3
iets 1 middel 2 vervoermiddel 1 voertuig 1 motorrijtuig 1 auto 1 bedrijfsauto 1
iets 1 middel 2 vervoermiddel 1 voertuig 1 motorrijtuig 1 auto 1 begrafenisauto 1
iets 1 middel 2 vervoermiddel 1 voertuig 1 motorrijtuig 1 auto 1 bezemwagen 1
iets 1 middel 2 vervoermiddel 1 voertuig 1 motorrijtuig 1 auto 1 brandweerauto 1
```

Incoherent collections of leaves, tops, cycles and mistakes at the higher levels (at the left side) can easily be detected in this way. After we made sure that large and crucial nodes in the Dutch hierarchy are acceptable, we replaced this chain by the WordNet1.5 equivalents if available. If there is not translation we maintained the Dutch word. The result for the above example looks as follows:

```
iets 1 means 1 conveyance 3 vehicle 1 automotive vehicle 1 auto 1 automatic 1
iets 1 means 1 conveyance 3 vehicle 1 automotive vehicle 1 auto 1 bedrijfsauto 1
iets 1 means 1 conveyance 3 vehicle 1 automotive vehicle 1 auto 1 hearse 1
iets 1 means 1 conveyance 3 vehicle 1 automotive vehicle 1 auto 1 bezemwagen 1
iets 1 means 1 conveyance 3 vehicle 1 automotive vehicle 1 auto 1 fire engine 1
```

If there are incoherences, these are due to wrong or bad translations because the Dutch hierarchy was already checked. Example of strange chain fragments are:

```
'chemical compound_1' --> 'building_1'
mind_7 -> group_1
```

All odd cases have been inspected and the translations have been corrected when necessary.

The translated chains have also been compared by FUE, but in that case the words have been replaced by the ILI-records. One comparison measured the number of ILI chains that are the same in all the wordnets, except for the fact that one of the wordnets missed an intermediate node. This comparison resulted, among others, in a list of nodes in the chains that have been covered by WordNet1.5, the Spanish wordnet and the Italian wordnet but are missing or different in the Dutch wordnet. We investigated the most important nodes in the other wordnets which are lacking in the Dutch wordnet and improved our subset where possible. The investigation concerned the nominal synsets covered by Spanish and Italian and not Dutch of the sub-chains of length 3, up to the frequency of 10.

The following explanations have been found for the gaps:

1. there is no equivalent:
  - 1.1. because it does not exist in Dutch: it is a genuine gap.
  - 1.2. it does exist but was not taken up in our resource:
    - 1.2.1. for non explicit reasons: just missed out, our out-dated
    - 1.2.2. because it would be a multi-word (which we did not include in our wordnet)
2. there is an equivalent:
  - 2.1. but we disagreed with the classification, therefore no changes were made
  - 2.2. we agree with the classification, but somehow it was not assigned (yet), so we adapted the classification.

One of the important conclusions from this comparison is that there are some major differences between the Dutch hierarchy and the WordNet1.5. hierarchy. Some important classes high up the hierarchy are treated differently, such as: "relation", "meaning", "communication", "thought", "area". This has important consequences for the comparison of the wordnets (see section 4 below). In the Appendix I some examples are given.

The above measures have lead to structural and quantitative improvements to the Dutch nominal hierarchy. The next table shows how the final nominal subset has been compiled.

*Table 1: Construction of the Dutch nominal wordnet*

<i>Selection criterion</i>	<i>Number of Synsets</i>
all synsets which have a manually processed relation (including the above improvements)	<b>9495</b>
all senses of Parole entries with a frequency above 100 and not part of the first Subset	<b>6005</b>
all synsets with only 1 or 2 automatically extracted equivalence relations	<b>7905</b>
all synsets with 10 or more relations and not included in the above union ( <b>18,376</b> )	<b>1554</b>
all direct hyponyms (1 level) of synsets with 50 or more relations	<b>9457</b>
all hyperonyms (all levels) of the above union ( <b>23,550</b> )	<b>5272</b>
Total	<b>24337</b>

The total set of 24,337 synsets corresponds with 40,535 word senses and 33,925 entries for nouns. All these nominal synsets are related to via hyponymy-links to a single top: "iets" (anything). Note that also the two verbal tops "gebeuren" (to happen) and "zijn" (to be) are linked to this top via a Cross-Part-of-Speech hyponymy. The pronoun "iets" can be used to substitute anything, including verbs and verb phrases. The complete wordnet can thus be accessed from this node.

### 2.1.2. Improving the equivalence relations

We have manually assigned equivalence relations to WordNet1.5 for all important concepts in the above selections. 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. The possible target translations have been weighted using several heuristics (see D014D015, Vossen et al 1998, and see below). A number of synsets in the original Dutch database did not receive a translation by this procedure, either because the entry was missing in the bilingual dictionary, or the translation could not be found in WordNet1.5:

Table 2: Synsets without translation in the Dutch database

	<i>Number of Synsets in Vlis</i>	<i>Number of Synsets without translation to WordNet1.5</i>	<i>%</i>
<b>nouns</b>	52359	23398	44,69%
<b>verbs</b>	9125	1060	11,62%

Table 2 shows that the result for verbs (11,62% not translated) are much better than for nouns (44% not translated). This is due to the fact that the nominal part contains more specialized vocabulary. To improve this matching we have applied two additional techniques.

According to the introduction to the Dutch-English dictionary, many English words that are directly taken over in Dutch without change of meaning and pronunciation, have been omitted (to save space, assuming that Dutch speakers are familiar with the word). It therefore makes sense to directly match the non-translated Dutch entries to the WordNet1.5 entries. The results are given in the next table:

Table 3: Directly matching entries between the Dutch database and WordNet1.5.

	<i>Entries in Vlis</i>	<i>Entries in WordNet1.5</i>	<i>Intersecting Entries</i>
<b>nouns</b>	63962	88200	3981
<b>verbs</b>	8822	14734	9

Inspection of this list showed that all the 9 verbal matches were wrong but the nominal intersection contained many good matches. By intersecting the direct nominal matches with the synsets without translation the nominal matching has been improved as follows:

Table 4: New nominal translations generated by direct matches with WordNet1.5

	<i>Synsets without translation</i>	<i>Senses without translations</i>	<i>Entries without translations</i>	<i>Overlap with WordNet1.5 noun entries</i>	<i>Matched synsets</i>	<i>Remaining unmatched synsets</i>
<b>nouns in the Van Dale database</b>	23398	27894	27053	841	726	22672

A second improvement consisted of reversing the English-Dutch dictionary into a Dutch-English dictionary, assuming that the set of Dutch translations is different from the set of Dutch entries. A third improvement consisted of varying the use of hyphens and spaces in the translations. In many cases, hyphens and spaces are used inconsistently, e.g. *animal park*, *animal-park* and *animalpark*. By replacing and removing spaces and hyphens in the translations we could further translate another 338 synsets. This resulted in the following improvements:

Table 5: Translations by Reversed English-Dutch dictionary and Replacing Spaces and Hyphens.

	Total Number of Synsets	Synsets without translation	Synsets matched by reversed English-Dutch dictionary	Synsets matched by replacing Spaces and Hyphens	Remaining unmatched synsets	% of total
<b>nouns</b>	52359	22672	2161	338	20180	38.54%
<b>verb</b>	9125	1060	183	7	869	9.52%

Since not all nouns from the Van Dale database are selected for the Dutch wordnet, the result for the wordnet, which is a subset of the complete database, is different:

Table 6: Synsets without translation in the Dutch wordnet

	Dutch WordNet	No ILI match	% of Dutch Wordnet
<b>nouns</b>	24337	4949	20.34%
<b>verbs</b>	9125	869	9.52%

For the Dutch wordnet, 80% of the selected noun synsets has a match to a WordNet1.5 synset. The fact that this is a better result than for the complete database has to do with the way synsets have been selected. First of all, we selected synsets with manual and reliable translations and, secondly, we excluded more specific levels that are more likely not translated (and probably cannot be translated).

In addition to increasing the coverage, we worked on improving the quality of the equivalence relations, along the following lines:

- Manual Improvement of unreliable Equivalent relations;
- Inspecting sorted flat trees after the nodes in the Dutch hierarchy have been replaced by their WordNet translations;
- Improving the Heuristics for automatically matching synsets.

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. 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. A large number of matches with a low score therefore indicates that the system had poor evidence for matching. 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 (the correct translation probably does not exist):

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

In general, we can thus state that many matches and/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

It often appeared that polysemous words with a badly translated sense also had poor translations for the other senses. We then manually translated all the senses of such a polysemous word. 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.

The next step was to improve the matching algorithm. As explained in D014D015 (Vossen et al. 1998), the matching algorithm, which is based on Agirre and Rigau (1996), tries to weight candidate translations by calculating the distance in the WordNet1.5 hierarchy of each translation to the translations of the Dutch hyponyms and hyperonyms. 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. Table 9a,b below show the results.

In addition we have implemented two new heuristics as well:

- 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 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:
  - 3.2.each time an English variant has the Dutch source as its translation
  - 3.3.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)<sup>3</sup>:

Dutch Vars:	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 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 the EWN Top Concepts (TCs) to the Base Concepts (BCs) in the Dutch wordnet and WordNet1.5. 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 target 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:

<sup>3</sup> In the real situation there are many more matches and translations, but we have listed here just two of them to illustrate the example.

hart 1	orgaan 1 ( <i>Living Part</i> ) deel 2 ( <i>Part</i> ) iets 1 LEAF
heart 1	playing card 1 card 1 ( <i>Artifact Function Object</i> ) paper 6 ( <i>Artifact Solid</i> ) material 5 ( <i>Substance</i> ) matter 1 inanimate object 1 entity 1 LEAF
heart 2	disposition 2 ( <i>Dynamic Experience Mental</i> ) nature 1 trait 1 ( <i>Property</i> ) attribute 1 ( <i>Property</i> ) abstraction 1 LEAF
heart 3	bravery 1 spirit 1 character 1 trait 1 ( <i>Property</i> ) attribute 1 ( <i>Property</i> ) abstraction 1 LEAF
heart 4	internal organ 1 organ 4 ( <i>Living Part</i> ) body part 1 ( <i>Living Part</i> ) 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 and the diversity of features. Currently, we have limited ourselves to the 63 features from the EuroWordNet top ontology (see Deliverable D017D034D036). 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 EWN ontology, and that most tops in the Dutch wordnet are unified to 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 EWN TCs (in other words have not been selected as Base Concepts):

Table 7: Synsets classified with top-concepts

	<i>Noun Synsets with Top Concepts</i>	<i>Verb Synset with Top Concepts</i>
<b>Dutch wordnet</b>	1170	836
<b>WordNet1.5</b>	793	617

Furthermore, we have converted the lexicographer's file codes in WordNet1.5 to make them compatible with the 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 8: Conversion of WordNet1.5 Lexicographer's file codes to EuroWordNet top-concepts

Code	WordNet File Name	EuroWordNet Top Concepts
03	noun.Tops	
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; Location; Dynamic;
35	verb.contact	Location; Dynamic;
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 9a and 9b 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 before the above measures have been taken (the Subset1 wordnet), 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 less than 5. 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 (all wrong) 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 for the first subset core wordnet, 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 subset results.

Table 9a: Automatic matching results for nouns

Nouns Correct Match	Subet1 Tree-matching before manual improv.		Tree-Matching after manual improv		Reversed translation		Top-Concept Matching		Gain Subet2
<b>1</b>	60	64,5%	70	66,6%	72	68,5%	74	70,4%	5,9%
<b>2</b>	8	8,6%	10	9,5%	8	7,6%	9	8,5%	-0,03%
<b>3</b>	4	4,3%	3	2,8%	3	2,8%	2	1,9%	-2,4%
<b>4</b>	0	0,0%	2	1,9%	3	2,8%	1	0,9%	0,9%
<b>5</b>	2	2,1%	2	1,9%	1	0,9%	1	0,9%	-1,2%
<b>&gt;5</b>	1	1,0%	1	0,9%	2	1,9%	2	1,9%	0,8%
<b>gap</b>	6	6,4%	9	8,5%	10	9,5%	10	9,5%	3,0%
<b>all wrong</b>	9	9,6%	3	2,8%	4	3,8%	4	3,8%	-5,8%
<b>hyper</b>	3	3,2%	5	4,7%	2	1,9%	2	1,9%	-1,3%
<b>Subtotal</b>	93		105		105		105		
<b>notrans</b>	102	52,3%	90	46,1%	90	46,1%	90	46,1%	-6,1%
<b>Total</b>	195		195		195		195		
<b>Top-3</b>	72	77,4%	83	79,0%	83	79,0%	85	80,9%	3,5%

Table 9b: Automatic matching results for verbs

Verbs Correct Match	Subet1 Tree- matching before manual improv		Tree-Matching after manual improv		Reversed translation		Top-Concept Matching		Gain Subet2
<b>1</b>	28	33,3%	28	32,9%	32	37,6%	39	45,8%	12,5%
<b>2</b>	7	8,3%	14	16,4%	18	21,1%	12	14,1%	5,7%
<b>3</b>	9	10,7%	10	11,7%	4	4,7%	5	5,8%	-4,8%
<b>4</b>	2	2,3%	3	3,5%	4	4,7%	5	5,8%	3,5%
<b>5</b>	1	1,1%	4	4,7%	3	3,5%	4	4,7%	3,5%
<b>&gt;5</b>	0	0,0%	6	7,0%	4	4,7%	2	2,3%	2,3%
<b>gap</b>	8	9,5%	10	11,7%	12	14,1%	10	11,7%	2,2%
<b>all wrong</b>	19	22,6%	6	7,0%	4	4,7%	4	4,7%	-17,9%
<b>hyper</b>	4	4,7%	4	4,7%	4	4,7%	4	4,7%	-0,06%
<b>Subtotal</b>	78		85		85		85		0,0%
<b>notrans</b>	6	7,1%	0	0,0%	0	0,0%	0		-7,1%
<b>Total</b>	84		85		85		85		
<b>Top-3</b>	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 subset1. 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: 12,5%. 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 changed as well due to the fact that more words have been translated by the measures explained previously.

For future work, the selections of nouns and verbs will remain fixed. All further efforts will focus on improving and extending the relations between the synsets. In the next section we will give the tables summarizing the data.

### 2.1.3. Overview tables for the Dutch wordnet (Subset2)

The next tables give an overview of the Dutch wordnet in its current state. In the near future, these figures will change, although the remainder work will focus on improving the quality rather than the quantity of data.

Table 10 below gives the basic quantitative statistics of the current Dutch wordnet. Compared to Subset1 for Dutch the major differences are:

- major increase in quantity of synsets (factor 5 for nouns and factor 3 for verbs)
- slight reduction of the number variants per synset
- slight increase of the polysemy
- major increase in the language internal relations (factor 3.5 for nouns and factor 2 for verbs)
- decrease of the average number of relations per synset
- major increase in the equivalence relations (factor 4 for nouns and factor 2.5 for verbs)
- slight decrease of the average number of equivalence relations per synset
- reasonable increase of the synsets without III link (factor 3 for nouns and factor 2 for verbs)

Table 10: Subset2 Overview NL

	Nouns	Verbs	Others	Total
Synsets	24337	9125	304	33766
Number of senses (variants)	40535	14278	857	55670
X variants per synset	1.67	1.56	2.82	1.65
Corresponding to number of entries (words)	33925	8868	785	43578
X senses per word	1.19	1.61	1.09	1.28
Language Internal Relations	56376	22815	402	79593
Average per synset	2.3	2.5	1.3	2.3
Equivalent Relations to ILI (WN1.5)	31237	14092	n.a.	45329
Average per synset	1.28	1.54	n.a.	1.34
Synset without ILI	5022	876	n.a.	5898

The figures first of all confirm the change in strategy for Subset2. Whereas we focussed on a rich and high-quality encoding of the core wordnets in Subset1, we now focussed on producing large quantities with relatively less rich encoding.

The current set thus contains more synsets and word meanings than required (50K word senses and 30K synsets) but as can be expected the average number of relations per synset has dropped. A good thing is the reduction of the variants per synset, which is due to a more critical encoding of synonymy. Polysemy increases as a result of the larger coverage. Finally, the lower average of the equivalence links indicates a better and more critical encoding as well. In the ideal case, there would be 1 equivalent for each synset.

Table 11 gives the distribution of the language-internal relations. As expected, the major increase is in the hyponymy relations. In the core wordnet (Subset1), about 75% of the relations was hyponymy, in the current wordnet 85%. Still, some of the other relations have increased as well: HAS\_HOLO/MERO\_PART (factor 2), INVOLVED/ROLE\_PATIENT (factor 2), IS\_CAUSED\_BY (factor 2), ROLE, INVOLVED/ROLE\_INSTRUMENT (factor 1,5). Note that some of the general relations, HAS\_HOLONYM and HAS\_MERONYM, have decreased because they have been differentiated into

more specific relations.

Table 11: Language Internal Relations NL

Language Internal Relations	Nouns	Verbs	Others	Total
Synsets	24337	9125	304	33766
BE_IN_STATE	136			136
CAUSES	252	743		995
HAS_HYPERONYM	24650	9809		34459
HAS_HYPONYM	24650	9809		34459
HAS_HOLONYM	188			188
HAS_HOLO_LOCATION	113			113
HAS_HOLO_MADEOF	131			131
HAS_HOLO_MEMBER	193			193
HAS_HOLO_PART	923			923
HAS_HOLO_PORTION	55			55
HAS_MERONYM	294			294
HAS_MERO_LOCATION	113			113
HAS_MERO_MADEOF	132			132
HAS_MERO_MEMBER	194			194
HAS_MERO_PART	926			926
HAS_MERO_PORTION	56			56
HAS_SUBEVENT	126	140		266
HAS_XPOS_HYPERONYM	22	44	5	71
HAS_XPOS_HYPONYM	51	18	1	70
INVOLVED	67	83		150
INVOLVED_AGENT	36	70		106
INVOLVED_DIRECTION		4		4
INVOLVED_INSTRUMENT	72	342		414
INVOLVED_LOCATION	29	34		63
INVOLVED_PATIENT	247	426		673
INVOLVED_SOURCE_DIRECTION	16	1		17
INVOLVED_TARGET_DIRECTION	2	21		23
IS_CAUSED_BY	351	325	355	1031
IS_SUBEVENT_OF	113	162		275
NEAR_ANTONYM	187	293		480
NEAR_SYNONYM	165	87		252
ROLE	151			151
ROLE_AGENT	104			104
ROLE_DIRECTION	5			5
ROLE_INSTRUMENT	408			408
ROLE_LOCATION	61			61
ROLE_PATIENT	664			664
ROLE_SOURCE_DIRECTION	16			16
ROLE_TARGET_DIRECTION	20			20
STATE_OF	25	7		32
XPOS_NEAR_ANTONYM	6	3		9
XPOS_NEAR_SYNONYM	426	394	41	861
Total	56376	22815	402	79593
Average per synset	2.32	2.50	1.32	2.36

The next table gives the number of relations taken over from the Van Dale database or added manually:

Table 12: Status of the Language Internal Relations NL

Language Internal Relations	Nouns	Verbs	Total	Percentages	
<b>Vlis &amp; Okay</b>	5771	2834	8605	28,20%	of Vlis Total
<b>Vlis &amp; ?</b>	16496	5418	21914	71,80%	of Vlis Total
<b>Vlis Total</b>	22267	8252	30519	67,83%	of All
<b>Manual &amp; Okay</b>	7794	4414	12208	84,36%	of manual Total
<b>Manual &amp; ?</b>	1877	387	2264	15,64%	of manual Total
<b>Manual Total</b>	9671	4801	14472	32,17%	of All
<b>Total</b>	31938	13053	44991		

Compared to the core wordnet, we can see that the contribution of the Van Dale relations has increased with 10%. The more specific levels are more reliable and we often could copy the relations as they are. We expect that from now on the manual encoding will only increase, because the total set is more or less stable and no relations are taken over from Van Dale.

Table 13 gives an overview of the different types of equivalence relations for the Dutch wordnet. Because of the manual encoding we see an increase in exotic types of equivalence relations. Especially, the number of EQ\_HAS\_HYPERONYM translations has increased. Many of these turn out to be (possible) gaps for which we could not find an appropriate translation by hand.

Table 13: Equivalence Relations NL

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

The next table gives the reliability of the translations based on the samples described in the previous section:

Table 14: Quality of the equivalence relations

Matching Type	Nouns			Verbs		
	No of synsets	Perc.	Reliability	No of Synsets	Perc.	Reliability
<b>manual/ok</b>	4138	17,00%	100%	3383	37,07%	100%
<b>1 match</b>	4846	19,91%	86%	763	8,36%	78%
<b>2 matches</b>	3059	12,57%	68%	652	7,15%	71%
<b>3-9 matches</b>	5408	22,22%	65%	2471	27,08%	49%
<b>10+ matches</b>	1864	7,66%	54%	980	10,74%	23%
<b>0 matches</b>	5022	20,64%	n.a.	876	9,60%	n.a.
<b>Total</b>	24337			9125		

Even though the total set of synsets has increased considerably, we still see that relatively-more synsets are translated by hand. For the core wordnets, 22% of noun translations and 7% of the verb translations was manual. In Subset2 the figures are 17% of the noun synsets and 37% (!!!) of the verb synsets are manually translated. Still, the manual translation of the verb synsets will be continued because the reliability of the poorly matched synsets (more than 3 matches) is too low: making up 37% of all the synsets. In the case of the nouns, we can rely more on the automatic techniques. Nevertheless, we will also continue to manually translate nominal synsets with more than 3 translations.

## 2.2. Subset2 for the Italian wordnet

In the second building phase for the Italian wordnet, we focussed on:

1. extending the wordnet to the full size of 50,000 senses (about 35,000 noun senses and 15,000 verb senses), mainly by adding missing Parole entries or analyzing the lexical gaps resulted from the comparison among the wordnets
2. improving synonymy relations
3. improving taxonomy consistency, especially for the second order hierarchies
4. adding relations which are necessary to precisely locate concepts which are not lexicalized in English
5. increasing the number of the equivalence relations.

Task 1 will be discussed in subsection 2.2.1; tasks 2, 3 and 4 will be discussed in 2.2.2; the work on the equivalence relations will be described in subsection 2.2.3. Finally, subsection 2.2.4. gives the overview tables for the Italian wordnet.

### 2.2.1. Increase of coverage

The first subset for Italian covered 18,934 nominal synsets and 3,692 verbal synsets. These corresponded to 19,646 noun senses and 4,577 verb senses. This subset was based on the common and local noun and verb Base Concepts, first level hyponyms of the BCs and for some taxonomies also other level hyponyms. Moreover, some synsets were encoded containing the adjectives/adverbs linked to the nouns and verbs by means of various internal relations. However, while for the nouns and verbs complete wordnets were built, mainly based on the hyponymy relation, with respect to the adjectival/adverbial synsets no hyponymy relations were encoded.

This core wordnet had to be extended to approximately 50,000 senses (about 35,000 noun senses and 15,000 verb senses). To extend the noun core subset, first of all we extracted from our main source those top-frequent entries of the Parole lexicon (about 5.000 entries) which had not already been included in our wordnet. These new entries were converted from the source to the Polaris import format and we performed a careful manual check of their hyperonyms associating each of them with the relevant entry in the Italian EWN database or, where necessary, inserting a new entry. For the verbs we extended our set to the complete lexicon made available by our source database. By manually checking information found within different sources, the verb senses distinguished have been grouped into synsets. Since, as will be explained below, the work on the verbs has required a huge manual effort, we are now starting to compare our Subset2 lexicon to the Parole lexicon, in order to (eventually) add top-frequent Parole entries missing in our source by manually creating new synsets/senses in our wordnet. Moreover, both for the nouns and the verbs we have also analyzed the lexical gaps resulted from the comparison of our wordnet to the other wordnets. When there was a gap the following situations generally occurred:

1. we had the word meaning in our wordnet but it had not been translated yet when the comparison was performed:  
e.g. {cellulose} corresponds to the Italian *cellulosa* which is encoded in our wordnet but did not have an eq\_link to the ILI when the comparison was performed.

2. we did not have the word meaning in our wordnet, because either it is not an Italian lexicalization, or it is simply missing in our source (although we have the concept in our language), or it is present in our source but it had not been inserted in our wordnet yet: e.g.
  - a) the {change\_state, turn} synset has no correspondent in our wordnet, because we only have a more general lexicalization, i.e. {cambiare, mutare, variare}, corresponding to the {change} ILI synset which is a hyperonym of {change\_state, turn}. In cases like this, we added an eq\_link between the concept lexicalized in Italian and the missing concept (in this particular case an eq\_has\_hyponym link from {cambiare, mutare, variare} to {change\_state, turn} was encoded);
  - b) the {enkindle, kindle} synset had no correspondent in our source database, although we have this concept lexicalized in Italian. In these cases we manually added the Italian synset corresponding to the ILI one;
  - c) the {china\_clay, china\_stone, kaolin, kaolinite, porcelain\_clay, terra\_alba, caolino} synset has a correspondent in Italian (*caolino*) which is present in our source but had not been encoded in the wordnet. This and other concepts have thus been added to the wordnet.
3. we have the word meaning but not in the same position within the wordnet. In these cases, either we restructured our data because we agreed with the classification in the other wordnets or we maintained our classification for various reasons: e.g.
  - a) in our source the synset for {entity} has not a high position because this word defines just a few entries (which are not often tops themselves), while in WN1.5 it has a very high position in the taxonomy. However, it seemed to us that the classification was not correct in our source, since in Italian there is a sense for the word *entità* which corresponds exactly to the WN sense and could be considered a very high top. Thus, we restructured our taxonomy by creating a very high top {entità};
  - b) the {stone, lapidate, kill\_by\_stoning} synset is a hyponym of {kill} in WN1.5, while the corresponding synset in Italian was a hyponym of {colpire} (to hit). In this case we decided to encode a multiple hyperonymy relation to both {colpire} and {uccidere} (to kill), since both the classifications seemed to be plausible.

### 2.2.2 *Building synsets, restructuring taxonomical chains and encoding additional relations*

For Italian the comparison with the Dutch and Spanish first subset wordnets had evidenced a lower synonymy density. The Italian database showed a very low ratio between the number of synsets and the number of variants, especially in the noun subset. This was due to the fact that the extraction of the first level hyponyms had not yet been followed by a systematic reorganization of those senses on the basis of synonymy relations. Indeed, in our main source we have an indication of synonymy between words (either in the form of synonymous definitions or with an explicit indication of the existence of such a relation) and we have also used an electronic dictionary containing only information on synonymy between words, however in both sources information on synonymy is ambiguous in that it is given either between entries or between a word sense and a word entry. Thus, we had to manually disambiguate the specific senses of the entries indicated as synonyms. A huge manual effort was therefore devoted at creating synsets in the intermediate levels of our taxonomies (synonymy is rare at the leaf level), and this work should still be refined.

Much work has then been devoted at restructuring our taxonomical chains. Our source contains very 'flat' hierarchies and we had to manually check many taxonomies in order to create more consistent hierarchies. As mentioned above, the analysis of the gaps resulted from the comparison with the other wordnets helped us to restructure our taxonomies, when:

- a) concepts were missing in our source but lexicalized in our language;
- b) concepts were not used as tops in our source but we realized they should be considered tops;
- c) concepts were classified differently in our source but we (more or less) agreed with classifications in the other wordnets.

In order to provide a more structured organization of the verbs, reflecting both the basic organization of the Top Ontology and the theoretical view behind it (according to which a verb sense basically refers either to a static or a dynamic situation), all the verb synsets in the Italian wordnet have been manually linked, by means of a (direct or indirect) hyponymy relation, either to the {essere} (to be) synset which is the top node indicating stativity, or to the {avere luogo, accadere, avvenire, occorrere...} (happen, take place) synset which is the top indicating non-stativity. These first two tops have below, as direct hyponyms, a number of other basic tops, identifying semantic classes of verbs which have been distinguished by using data found within our source database, but also by taking into account the results of research in the theoretical field. For instance, *diventare* (to become) and *rendere* (to make) have no hyperonyms in our source, however the former has been linked, by means of a has\_hyperonym relation, to *cambiare* (to change) and the latter to *causare* (to cause). These two tops have been directly linked to the *accadere* synset. This re-organization of our taxonomies was done because *diventare* is one of the inchoative hyperonyms found in our source and *rendere* one of the causative hyperonyms. By grouping them under the relevant tops, we may account for similarities of meaning between each of their taxonomies and other inchoative or causative taxonomies found in the database; this may have important consequences with respect to the future use of the database, since inchoativity/causativity have been demonstrated to be linked both to other semantic properties (different *Aktionsarten*) and to different syntactic properties.

As already explained in deliverable D014015 (Vossen et al. 1998), in our source database relations for verbs, included hyponymy relation, had already been made explicit only for some taxonomies/groups of words, thus we had to perform a manual sense disambiguation of most of verb hyperonyms. With respect to this task the work on the second subset required less effort than the work on the first subset, due to the fact that we encoded data on verbs at the lowest levels within taxonomies. These verb hyperonyms are generally less polysemous than BCs, thus disambiguating the hyperonym senses at this stage was not as difficult as for the BC first level hyponyms. However, apart from hyponymy various other links have been encoded (see table in section 2.2.4.), both for the verbs and the nouns, mainly by manually adding relations indicated within definitions. Since words at low levels within taxonomies have generally very specific senses and typical Italian lexicalizations are found here (which cannot be properly translated into WordNet), we worked to add more detailed information on their meanings by encoding richer sets of relations in order to precisely locate these concepts in the net. For instance, *rincasare* (to go back home) is a hyponym of *tornare* (go back) which has only three senses in our database, one of which is very rare. Thus, it was not very difficult to disambiguate the sense of *tornare* involved within the definition of *rincasare*, i.e. sense 1: "go back to the place where one left from or went away from." *Rincasare*, however, is a typical lexicalization found in Italian but not in English, and a synset corresponding to it is not found in WN1.5. Thus, it was necessary to add relations indicating the particular meaning components involved within the meaning of the verb: in this case an 'involved\_target\_direction' relation with the noun *casa* (home) has been added.

### 2.2.3. Encoding equivalence relations

We have manually assigned equivalence relations to WN1.5 for all important concepts in our selections. The remaining synsets have been first translated into English by using the bilingual Italian-English Collins dictionary and these translations have then been mapped to WN1.5, using as far as possible very simple semi-automatic procedures. A part of the Italian synsets could not be mapped because either the entries were missing in the bilingual dictionary or possible translations for them could not be found in WN1.5:

Table 15: Synsets not mapped to WN1.5

	<i>Number of Entries missing in the bilingual dictionary</i>	<i>%</i>	<i>Number of Synsets without translation to WordNet1.5</i>	<i>%</i>
<b>nouns</b>	4495	22%	898	3.4%
<b>verbs</b>	1164	18%	309	3.5%

The semi-automatic mapping procedure we first used worked comparing our taxonomies with the WN1.5 ones. First of all a group of Italian nominal Base Concepts was circumscribed and, starting from these pivot-points already mapped, the software performed the vertical extraction of the first level hyponyms. At the same time, all the ILI records of the corresponding taxonomy were extracted. Then a translation from Italian to English was automatically performed and compared with the result of the extraction from WN1.5: thus `eq_synonymy` or `eq_near_synonymy` was automatically stated when a matching was found within corresponding taxonomies in the two wordnets. Unfortunately, the average of analysed entries which were successfully mapped by means of an `eq_synonymy` or `eq_near_synonymy` relation by using this procedure was, on the whole, rather low. Apart from the high number of entries not found in the bilingual, the effectiveness of the mapping procedure varied depending on the different kind of noun classes it analysed. The results, which were satisfactory in the case of concrete nouns such as *animals*, *plants*, the more common *instruments* and *vehicles*, were not fine with abstract nouns such as *act*, *event*, *process*, *phenomenon*. In our source, the hyponyms of these Base concepts are listed for the most part under the *act/effect* taxonomies which are very large and undifferentiated. This kind of second-order nouns gave us problems when trying to automatically map them to the ILI using this procedure because of the differences of classification when compared with WN1.5. Thus, in many cases, no automatic mapping was possible.

Due to these problems with `second_order` entities, we tried to refine our procedure adding further heuristics which could help to semi-automatically encode a higher number of `eq_relations` both for the nouns and the verbs. Thus, first of all we decided to add a score to the output, automatically assigned by taking into consideration the levels within the taxonomy in which the Italian and English mapping concepts are found. E.g., if they have the same direct hyperonym the relation encoded obtains a very high score; the score decreases, instead, for each hyperonym which is not shared by the two concepts and occurs under the shared hyperonym. As said above, we noticed that taxonomies of second-order entities in our database are often built differently with respect to WN1.5. This seems to be due both to the nature of second-order word meanings and to the way in which WN on one hand and our database on the other are built. Whereas WN contains very deep hierarchies, with very granular sense distinctions, our database is well structured only for some taxonomies (which we have manually restructured), while most of the hierarchies are rather flat and there is not the over-differentiation of meanings found in WN. This results in a generally different structure, where hyperonymy is often stated differently. Thus, we decided that if the synset/s containing the translation of an Italian word is/are not found in WN1.5 within the same taxonomy

of the Italian word synset, but the procedure still finds one or more synsets corresponding to ours, these synsets have to be returned as possible `eq_near_synonyms` together with a score determined by simply taking into account the number of possible translations found: the more translations are found the lower is the score. We have then started to manually revise the automatically produced mappings, checking both relations with very low scores and relations with higher scores in order to precisely evaluate the degree of reliability of the different outputs. Thus, by analysing a subset of all the `eq_relations` automatically encoded for words not occurring within the same taxonomies of the corresponding concepts in WN, we saw that when one only `eq_relation` is stated, in 96% of the cases analysed this is correct and mostly an `eq_synonymy` relation. When two `eq_relations` are automatically stated, we have about 47% of the cases in which `eq_near_synonymy` relations with the two synsets found is correctly stated; 38% of the cases in which one of the two `eq_relations` is correct; and 15% of the cases in which both equivalences are wrong. Of course, in cases in which three or more mappings are found the situation is still more various and we find many errors.

Since we feel that `eq_relations` need to be sufficiently reliable to make our data re-usable in applications, and due to all the (semi-automatic and manual) work we had to carry out also to encode `internal_links` and to structure our wordnet, instead of trying to envisage and develop more complex heuristics which could in any case hardly produce a very high percentage of correct results, we decided to manually check the whole output of our procedure, and in particular to revise all the problematic cases, i.e. cases with lower scores. That is:

- we verify if the automatically stated mappings to concepts occurring within the same taxonomies in WN1.5 are in any case correct;
- we verify if the mappings stated to the only one candidate concept found in WN1.5 are also correct;
- all the other cases are being manually carefully revised, starting from those with the lowest scores;
- we are also checking the entries which haven't been translated into English by using the bilingual, and in case we find important concepts we manually add the translations.

Thus, the equivalent relations encoded within the database so far have been mapped with high accuracy, since they have been all manually checked or selected from the output of the automatic procedure. The remaining equivalent relations obtained by using the procedure are all being manually revised and will only be added to the wordnet after this revision. In the following table the number of relations encoded so far in the Italian wordnet in each of the ways seen is indicated:

*Table 16: Synsets mapped using different methodologies*

	<i>Number of equivalent relations encoded so far</i>	<i>Number of synsets automatically mapped and manually checked</i>	<i>%</i>	<i>Number of synsets manually mapped</i>	<i>%</i>
<b>nouns</b>	12176	7000	57%	5176	42.5%
<b>verbs</b>	3266	872	26.7%	2394	73.3%

### 2.2.3. Overview tables for the Italian wordnet (Subsets1+2)

Table 17: Italian Subsets1+2 Overview

	<i>Nouns</i>	<i>Verbs</i>	<i>Others</i>	<i>Total</i>
<b>Synsets</b>	26466	8805	1822	37093
<b>Number of senses (variants)</b>	31318	11847	1829	44994
<b>X variants per synset</b>	1.2	1.34	1	1.21
<b>Corresponding to number of entries (words)</b>	20434	6445	1829	28708
<b>X senses per word</b>	1.3	1.83	1	1.56
<b>Language Internal Relations</b>	66465	28463	2037	96965
<b>Average per synset</b>	2.51	3.23	1.11	2.6
<b>Equivalent Relations to ILI (WN1.5)</b>	12176	3266	-	15442
<b>Average per synset</b>	0.5	0.37	-	0.43
<b>Synset without ILI</b>	14855	5539	-	20394

Table 18: Italian Language Internal Relations

Language Internal Relations	Nouns	Verbs	Others	Total
Synsets	26466	8805	1822	37093
BE_IN_STATE	123	20		143
CAUSES		660		660
HAS_HYPERONYM	26400	9122		35522
HAS_HYPONYM	26400	9122		35522
HAS_HOLONYM	266			266
HAS_HOLO_LOCATION	11			11
HAS_HOLO_MADEOF	162			162
HAS_HOLO_MEMBER	196			196
HAS_HOLO_PART	489			489
HAS_HOLO_PORTION				
HAS_MERONYM	266			266
HAS_MERO_LOCATION	11			11
HAS_MERO_MADEOF	162			162
HAS_MERO_MEMBER	196			196
HAS_MERO_PART	489			489
HAS_MERO_PORTION				
HAS_SUBEVENT		149		149
HAS_XPOS_HYPERONYM				
HAS_XPOS_HYPONYM				
INVOLVED	11	1271		1271
INVOLVED_AGENT	101	1155		1256
INVOLVED_DIRECTION		19		19
INVOLVED_INSTRUMENT	8	271		279
INVOLVED_LOCATION	33	68		102
INVOLVED_PATIENT	15	278		293
INVOLVED_SOURCE_DIRECTION		60		60
INVOLVED_TARGET_DIRECTION		27		27
INVOLVED_RESULT		83		83
IN_MANNER		58		58
IS_CAUSED_BY		201	459	660
IS_SUBEVENT_OF	17	132		149
NEAR_ANTONYM	30	27		57
NEAR_SYNONYM	486	10		496
ROLE	1282			1282
ROLE_AGENT	1256			1256
ROLE_DIRECTION	19			19
ROLE_INSTRUMENT	279			279
ROLE_LOCATION	102			102
ROLE_PATIENT	293			293
ROLE_SOURCE_DIRECTION	58		2	60
ROLE_TARGET_DIRECTION	21		6	27
ROLE_RESULT	83			83
IS_MANNER_FOR			58	58
STATE_OF	1		143	144
XPOS_NEAR_ANTONYM				
XPOS_NEAR_SYNONYM	7199	5830	1369	14398
Total	66465	28463	2037	96965
Average per synset	2.51	3.23	1.11	2.61

*Table 19: Italian Equivalence Relations\**

<i>Equivalence Relations</i>	<i>Nouns</i>	<i>Verbs</i>	<i>Total</i>
<b>EQ_HAS_HYPERONYM</b>	3314	583	3897
<b>EQ_HAS_HYPONYM</b>	13	12	25
<b>EQ_NEAR_SYNONYM</b>	1595	1898	2493
<b>EQ_SYNONYM</b>	7225	773	7998
<b>EQ_HOLONYM</b>	2		2
<b>EQ_MERONYM</b>	8		8
<b>EQ_INVOLVED</b>	7		7
<b>EQ_BE_IN_STATE</b>	1		1
<b>EQ_IS_CAUSED_BY</b>	11		11
<b>Total</b>	12176	3266	15442

\* As said above, all our synsets have been automatically mapped but, since we are manually revising all the mappings, here we indicate only the relations already manually revised and added to the wordnet.

### 2.3. Subset2 for the Spanish wordnet

In the first building phase, mainly because of the extensive use of automatic methods of lexical knowledge acquisition (cf. Atserias et al. 1997), FUE already got its subset practically to the full size which was aimed by the project (cf. Vossen et al. 1998a). Therefore, in this second phase we have focused on a mostly manual effort oriented to achieve the following two tasks:

1. improving the overlap across wordnets
2. enhancing the quality of the Spanish wordnet (henceforth SpWN)

Moreover, as a collateral effect, both tasks have also led to a small extension in quantity of the SpWN.

The analysis of the problem led FUE to adopt a strategy based on five inter-related actions. Such strategy is schematised in the flow chart shown by Figure 1 below.

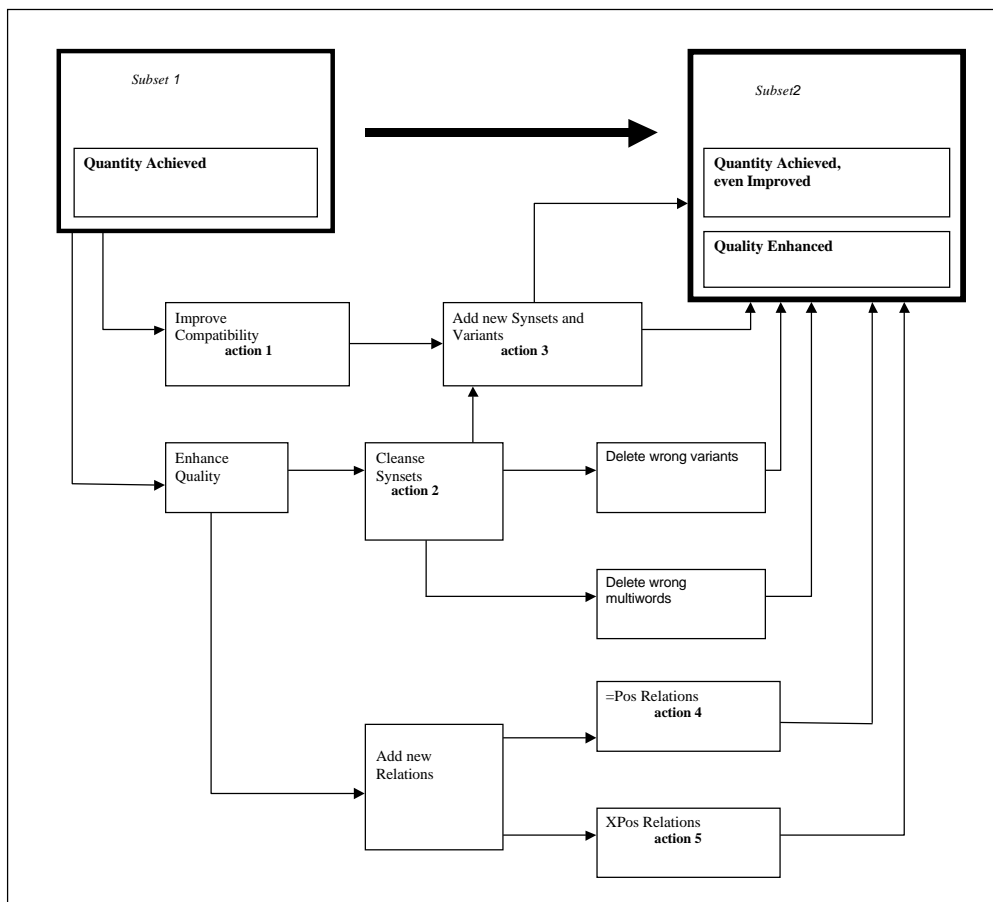


Figure 1. FUE Strategy for Subset2

Those five actions have been the following:

**Action 1:** Detection of gaps in hyponymy chains whose filling will cause a substantial increase in the compatibility between wordnets. Manual creation of such synsets.

**Action 2:** Manual cleansing of synsets containing automatically-generated variants. This

action has been based on two lists: (i) a list of pairs of nominal synsets which are adjacent in the hyponymy chain and share at least one automatically-generated variant; and (ii), the list of the multi-word expressions present in synsets. This action typically resulted in deletion of wrong variants from synsets, re-location of some of those variants in other pre-existing synsets, and creation of new synsets to re-locate the rest of such variants.

**Action 3:** Manual addition of new vocabulary which has been considered relevant — entailing the corresponding creation of new synsets. Such vocabulary mostly came from (i) a list of nominal synsets covered by the Catalan wordnet<sup>4</sup> but not by SpWN, and (ii) a working database of verbal synsets containing automatically-generated variants; but also (iii) from other small areas of vocabulary detected *incidentally* while carrying the effort due to any of the rest of actions.

**Action 4:** Addition of new relations between nominal synsets different from hyponymy and synonymy. This action was carried by (i) semi-automatic creation of *near-synonym* and *cause* relations imported from WN 1.5; and (ii) manual creation of other types of relations between synsets when the possibility was detected while carrying another action.

**Action 5:** Manual addition of *cross-part of speech* relations between nominal and verbal synsets. This work has been based on a list of noun-verb pairs obtained by means of morphological criteria.

The work involving operations on the content of synsets (Actions 1 to 3) is described in Section 2.3.1. Section 2.3.2 discusses the creation of relations between synsets. Last, Section 2.3.3. provides the overview tables for Subset 2.

### 2.3.1. Improving the quality of Subset 2: work on the content of synsets

As told above and largely discussed in Vossen et al. (1998a) and Rodríguez et al. (1998), the bulk of the SpWN as presented in Subset1 was created using the automatic methods of knowledge acquisition described in Atserias et al. (1997). The effects were that, on the one hand, the figures of synsets, senses and entries aimed at EWN were already achieved for Spanish in that phase; but, on the other hand, only about 25% of the variants had undergone manual creation or revision —the rest bearing an *index of confidence* in the automatic method which had generated them ranging from 85 to 97%. Another effect was that semantic relations standing between concepts were limited in type.

Consequently, we decided to act in order to improve the quality of the SpWN, in two directions, (i) manual revision of automatically-generated variants and (ii) creation of relations —specially those from types missing in Subset 1. Moreover, the simultaneous line of development common to all partners which consisted in directing efforts to improve compatibility between wordnets, was a task which naturally came to strengthen and complement this goal.

**Action1.** We are not going to describe here the full process of comparison of wordnets, since Section 3 below is devoted to it. It's just worth to be mentioned that preliminary comparisons, oriented to enhance compatibility between wordnets, led for every language, in the case which matters here, to a list of ILI-Chains with *gaps*, i.e. synsets covered by other languages but not for that of reference. Those *gaps* were classified according to the number of ILI-Chains in which they

---

<sup>4</sup> The Catalan wordnet is being developed within the framework of the project ITEM (TIC96-1243-C03-03), funded by the Spanish government to develop a series of linguistic resources —among which wordnets for Catalan and Basque (two languages spoken in Spain) to be compatible with the SpWN as developed in EuroWordNet. The autonomous Catalan government also funds the Catalan wordnet by its action CREL.

were involved. In the case of Spanish, the classification ranged increasingly from 1 to 52. This means that whether a gap rated 10 is filled in the wordnet of reference, automatically ten ILI-Chains (from top to bottom of the hierarchy) result to be complete, thus improving the compactness of the local wordnet and its compatibility to other language's.

In the case of Spanish *gaps* classified above 5 were filled, i.e. the corresponding synsets, senses and entries were created, amounting to about 200 synsets and 400 variants.

It must be noticed that this action, apart from improving the quality of the local wordnet, obviously entailed an increase in the quantity of word senses and entries which, joint to other increments described in **action 3**, came to act as a balance for the expected decrease in quantity caused by the process of synset cleansing which will be described next in **action 2**.

**Action 2.** Although the core of the nominal SpWN (5041 word senses) was built manually, the rest of nouns in Subset 1 were generated by high-confidence heuristics relying in translations provided by bilingual dictionaries. Grossly, such a procedure picked the appropriate translations for variants present in WN 1.5 synsets thus either creating or increasing an equivalent Spanish synset. It steadily became apparent that those variants needed to be manually inspected in order to achieve a higher quality wordnet.

Different is the case of verbs. All verbal synsets in Subset 1 came from a manual revision of a working-database consisting of automatically-generated synsets, amounting to 6795 word senses. Therefore, verbs in Subset 1 don't need further revision —just extension until the complete exhaustion of the working-database (see **action 3**).

Coming back to nouns, it became obvious that the task of verifying by hand the bulk of the 34579 automatically-generated variants, which showed a *confidence index* between 85 and 97%, should faced in some structured manner.

The study of the bilingual-dictionary-based approach followed to build Subset 1 showed two main sources of error: lack of precision in translation and paraphrases.

First, due to either differences in lexicalisation between languages or to the very structure of bilingual dictionaries, in many cases a word sense in the source language is translated by means of a more general lexicalized concept (a hypernym) in the target language, or/and by means of a list of more particular lexical concepts (a list of hyponyms). The repercussion in FUE's strategy for building the nominal part of Subset 1 is that, in many cases, a synset and its immediate hyponym both contain the one same variant. For that reason we generated a list of adjacent pairs of synsets showing this kind of repetitions which will serve as a guide and basis for manual inspection.

Second, in other cases, the lack of a lexicalized equivalent of the concept leads lexicographers to translate a word sense in the source language by means of a multi-word expression, which in the best cases correspond to an idiom which can be considered a (complex) lexical unit in the target language, but in the worst cases it is nothing but a monolingual-style definition. The repercussion in the SpWN is that some synsets bear multi-word candidates to variants which need to be inspected. Therefore, we generated such a list of synsets.

The list of multi-words in Subset 1 amounted to 3231. This list was pruned until only 887 genuine lexicalized idioms remain. The rest were deleted, thus causing in many cases the corresponding restructuring of the hierarchy.

The list of repetition-pairs is still under inspection. It is impossible to assess how many words, senses and synsets have been treated by this procedure. Every single intervention on one synset unchains a series of rebounding effects: variants can be just deleted but more usually they are moved to other synsets which at their turn happen to be a new cause for either revision or re-structuring of other related synsets; at this respect it has been estimated that each intervention on one synset cause a re-structuring of at least other five synsets. On the other hand, for reasons of saving time while revising synsets, we couldn't keep track of the number of all synsets and variants which after inspection were considered correct —so no action was carried on them.

Partial data from September 1998 have shown that at least 5555 noun word senses had been treated by this procedure —but, as explained above, this figure is necessarily lower than the real one. Nevertheless, the substantial increase in the quality of Subset 2 for nouns can be sensed from the fact that, comparing the summarising tables provided in section 2.3.3 with those for Subset 1 (Vossen et al. 1998a), in spite of the fact that Subset 2 has decreased its number of noun variants in 1510 with respect to Subset 1, its number of 100% confidence variants has increased in (at least) 2778.

**Action 3.** Subset 1 reached the borderline of the quantitative amount of data aimed at EWN but **Action 2** described above was expected to cause a reduction in the number of word senses, therefore it was considered necessary to carry some parallel actions to increase the number of synsets and senses to keep the quantity balanced.

The pre-requisite for these actions was that they should be performed manually, in order to also preserve/improve quality.

The first decision was to carry on loading verbs on the wordnet. Creation of verbal synsets has been always manual, even in Subset 1, on the basis of an automatically-generated working-database which is treated by linguists. This work has continued in subset 2 leading to an increment of 932 synsets and 1599 word senses.

The second line of development intended to increase relevant vocabulary involved nouns. This line is related to another project in which our group takes part —the building of a wordnet for Catalan. Comparing to the SpWN the basic vocabulary which has already been covered for Catalan a list of *gaps* in SpWN has been generated. Such a list is being treated manually in order to add to SpWN new synsets and senses belonging to concepts which, being relevant for Catalan and considering the practical cultural identity between speakers of both languages (actually, any speaker of Catalan is also a speaker of Spanish), necessarily they are relevant for Spanish. No specific quantification has been assessed for this action.

Last, it must be noticed that some micro-areas of vocabulary have been inserted *in passing* as linguists detected the need while carrying any of the rest of actions, either because they have been considered relevant enough but they were still missing, or because they were necessary to serve as grounds for some of the semantic relations which are described in the next section.

### 2.3.2. Improving the quality of Subset 2: creation of new relations between synsets

The second main task which is addressed to improve the quality of the wordnet involved establishing new relations between synsets. In Subset 1 the SpWN only included relations which could be directly imported from WN 1.5 because the work in that phase focused mainly in automatic methods of acquisition. In Subset 2 two actions have been addressed to enhance such a point —**action 4**, relations between synsets of the same morphosyntactic category; and **action 5** *cross-part-of-speech* relations.

**Action 4.** This action was performed in two directions. First, *cause* relations between verbs and *antonym* relations between nouns and between verbs (synsets) were automatically derived from WN 1.5 on the hypothesis that they will also hold for their equivalent synonym synsets in Spanish —when existing. The list was manually inspected, thus confirming the hypothesis. Therefore, in one case the corresponding *cause* relations were encoded; and in the other WN 1.5's *antonym* relations were encoded as *near antonym* relations in EWN. The reason for the latter is that *antonymy* in EWN holds between variants while this relation can only be extracted from WN 1.5 relating to synsets —equivalent to EWN's *near antonym* relation. Second, a number of other relations of different kinds were encoded directly by hand when finding the possibility as linguists found the possibility while carrying other actions.

#### **Action 5.**

In order to establish noun-verb relations, we have first produced automatically noun-verb pairs of words potentially related, and then manually established all the semantic relations between the synsets for the noun and the synsets for the verb.

To produce the candidate pairs there are different possibilities:

- Morphological derivation of nouns from the verbal stem, or the other way round, in Spanish.
- Search for patterns in dictionary definitions. For instance, "alicatador" defined as "Instrumento para alicatar" indicates an INVOLVED\_INSTRUMENT relation between some of the meanings of "alicatador" and "alicatar".
- Usage of the EuroWordNet InterLingual Index to extrapolate relations from other monolingual wordnets included in the database.

For the subset 2 of EuroWordNet we have concentrated on the first criteria. In particular, we have studied nouns derived morphologically from verbs. The data available was subset 1, with 3086 verb forms (for 7953 senses grouped into 3294 synsets). Discarding reflexive and multi-word verbs, we had 2231 verb forms to apply morphological derivations. The network of nouns had 23216 word forms, 41292 senses and 18577 synsets.

For every verb, we have automatically generated nouns according to the following suffixes and allomorphs: -ante, -dero, -dor, -ero, -torio,-ción, -mento, -miento, -imiento, -ón. Filtering out the nouns that were not present in the database, we obtained a list of 1271 candidate pairs. The list was divided into non-ambiguous pairs (when both the noun and the verb had only one possible meaning in the database) and ambiguous pairs (otherwise). Both lists were manually revised to identify relations. When there was a semantic relation but any of the adequate senses was not present in the database, it was introduced manually.

The results can be seen in Table 20. A total of 1623 noun-verb relations have been extracted (3246 in both directions). Morphological derivations has shown to be very productive, with 1.3 relations per candidate pair to obtain 0.5 relations per verbal synset, which seems a high degree of interconnectivity between the nouns and verbs networks of the Spanish wordnet.

An 87% of the candidate pairs produced noun-verb semantic relations. The other 13% can be grouped according to three different reasons:

- The presence of false derivatives, as in "turbante/turbar" ("turban/disturb").
- Derivatives with an ethymological relation that does not hold in actual use of the language, such as "restaurante/restaurar" ("restaurant/restore").
- Words belonging to the same family but only indirectly related, such as "aceitero/aceitar" ("oil bottle/to oil", which should both be related to "oil").

This experience let us evaluate the possibility of establishing noun-verb relations automatically according to morphological cues. Unfortunately, it seems a very hard task:

- 10% of the senses related had to be manually included to the database while checking the relations. This percentage of missing senses would not be detected by an automatic procedure.
- Even if the related senses are identified, choosing the appropriate semantic relation cannot be done only on the basis of the morphological cues. As it can be seen in Table 20, none of the suffixes indicates clearly one preferred semantic relation. For instance, -dor produces 460 ROLE\_AGENT and 224 ROLE\_INSTRUMENT, but also 19 ROLE\_LOCATION, 18 ROLE, and 12 XPOS\_FUZZYNYM.
- The absence of Spanish glosses produces a lack of contextual information to restrict possible relations.

Table 20: Spanish Relations derived from Suffixes

Suffix	Role	Agent	Patient	Instru	Loc	Result	XpFuz	Xpnsyn	Iscausb	total
-ante	35	126	1	22	-	-	2	1	-	187
-dero	12	2	4	47	55	-	9	-	-	129
-dor	18	460	1	224	19	-	12	-	-	735
-ero	1	22	-	4	-	-	12	-	-	39
-torio	4	-	-	4	15	-	-	3	-	26
-ción	4	1	-	-	-	1	-	105	3	114
-mento	11	-	-	3	-	-	-	1	-	15
-miento	32	1	-	3	-	11	-	216	6	269
-imiento	4	39	-	19	-	-	2	45	-	109
total	121	651	6	326	89	12	37	371	9	1623

Figures on table 20 are slightly different to those in table 22 due to interactions with the other actions.

### 2.3.3 Overview tables for the Spanish wordnet (Subset 2)

Tables below show an overview of Subset 2 as it is now. Nevertheless, it must be noticed that actions 2, 3 and 4 are still being carried on at the moment, so further improvement is still expected. In this section we will comment such tables by comparing them with those for Subset 1.

Table 21 shows that although this phase focused mainly on improving the quality of the wordnet, a task which in part involved deletion of variants —and even deletion of synsets, as in the case of non-lexicalized multiwords (see **action 2**)—, such a procedure was balanced by addition of new vocabulary (see **actions 1, 2 and 3**). What's more, Subset 2 even shows a small increase in quantity with respect to Subset 1: 2022 synsets, 89 variants and 711 word entries more.

By the way, the substantial difference in increase between synsets and variants, suggests that a relatively stable amount of variants has been re-distributed in a larger number of synsets, thus showing that the qualitative structure of the wordnet has been improved.

Other data which point in this direction are the decrease of variants per synset and senses per word in Subset 2: 2.08 vps and 1.84 spw, compared to 2.27 vps and 1.88 spw in Subset 1. Such figures suggest that some of the artificial polysemy which Subset 1 could bear has been reduced.

Now looking at relations, table 21 shows an increase in 8260 relations with respect to Subset 1. Remarkably 2661 of those relations are *cross-part-of-speech* —zero relations of this kind in Subset 1.

A more precise overview on this point can be seen in table 22. It shows how the most substantial increase in relations different from hyperonymy and hyponymy comes from noun-verb relations —*involved*, *role* and *XPOS\_near\_synonym*. Other increments with respect to Subset 1 are: holonyms and meronyms (1760 relations more in Subset 2), *near\_antonym* (825 in Subset 2 vs. 0 in Subset 1), and remarkably, hyperonymy-hyponymy for verbs (2624 relations more than in Subset 1) which correspond to the continuing effort in encoding verbal synsets (see **action 3**).

Equivalence relations stay relatively stable with respect to Subset 1. The small increment shown by table 23, 1832 relations more, is practically equivalent to the increase in number of synsets.

With respect to this point, it must be remarked that the total amount of 23201 synsets correspond to 23068 equivalence relations to ILI. Coming back to table 21, we can see that only 189 synsets are not linked to the interlingua. Therefore we can consider that the cross-linguistic compatibility of the Spanish wordnet is practically complete (99.18%)

Last, Table 24 shows a remarkable increase in manually encoded variants —4377 more than Subset 1. In other terms, 33.7% of the variants in Subset 2 are manual. With respect to this, it must be remarked that, as explained above, this figure should be understood properly as that *at least* 33.7% of the variants have been manually encoded or revised, since we could not keep record of all those synsets which have been revised, found correct and therefore left unchanged.

Summing up, the Spanish wordnet as it is now in Subset 2, shows a complete equivalence to ILI, a substantial increase in quality, and a small increase in quantity, thus practically achieving the goals aimed at EWN.

Table 21 : Subset2 Overview ES

	<i>Nouns</i>	<i>Verbs</i>	<i>Others</i>	<i>Total</i>
<b>Synsets</b>	19663	3538	0	23201
<b>number of senses (variants)</b>	39782	8394	0	48176
<b>X variants per synset</b>	2.02	2.37	0	2.08
<b>Corresponding to number of entries (words)</b>	22881	3324	0	26205
<b>X senses per word</b>	1.74	2.53	0	1.84
<b>Language Internal Relations</b>	43151	6756	2661*	52568
<b>Average per synset</b>	2.19	1.91	?	2.27
<b>Equivalent Relations to ILI (WN1.5)</b>	19534	3534	0	23068
<b>Average per synset</b>	0.99	1.00	0	0.99
<b>Synset without ILI</b>	185	4	0	189
<b>Percentage of Synsets without translation</b>	1%	0%	0	1%

• These 2661 relations hold between nouns and verbs.

Table 22: Language Internal Relations ES

Language Internal Relations	Nouns	Verbs	Others	Total
BE_IN_STATE	0	0	0	0
CAUSES	0	97	25	122
HAS_HYPERONYM	19739	3142	0	22881
HAS_HYPONYM	19739	3142	0	22881
HAS_HOLONYM	0	0	0	0
HAS_HOLO_LOCATION	0	0	0	0
HAS_HOLO_MADEOF	89	0	0	89
HAS_HOLO_MEMBER	217	0	0	217
HAS_HOLO_PART	1257	0	0	1257
HAS_HOLO_PORTION	0	0	0	0
HAS_MERONYM	0	0	0	0
HAS_MERO_LOCATION	0	0	0	0
HAS_MERO_MADEOF	89	0	0	89
HAS_MERO_MEMBER	217	0	0	217
HAS_MERO_PART	1257	0	0	1257
HAS_MERO_PORTION	0	0	0	0
HAS_SUBEVENT	0	0	0	0
HAS_XPOS_HYPERONYM	0	0	0	0
HAS_XPOS_HYPONYM	0	0	0	0
INVOLVED	0	0	122	122
INVOLVED_AGENT	0	0	576	576
INVOLVED_DIRECTION	0	0	0	0
INVOLVED_INSTRUMENT	0	0	301	301
INVOLVED_LOCATION	0	0	86	86
INVOLVED_PATIENT	0	0	6	6
INVOLVED_SOURCE_DIRECTION	0	0	0	0
INVOLVED_TARGET_DIRECTION	0	0	0	0
IS_CAUSED_BY	0	97	25	122
IS_SUBEVENT_OF	0	0	0	0
NEAR_ANTONYM	547	278	0	825
NEAR_SYNONYM	0	0	0	0
ROLE	0	0	122	122
ROLE_AGENT	0	0	576	576
ROLE_DIRECTION	0	0	0	0
ROLE_INSTRUMENT	0	0	301	301
ROLE_LOCATION	0	0	86	86
ROLE_PATIENT	0	0	6	6
ROLE_SOURCE_DIRECTION	0	0	0	0
ROLE_TARGET_DIRECTION	0	0	0	0
STATE_OF	0	0	0	0
XPOS_FUZZYNYM	0	0	37	37
XPOS_NEAR_ANTONYM	0	0	0	0
XPOS_NEAR_SYNONYM	0	0	392	392
<b>Total</b>	<b>43151</b>	<b>6756</b>	<b>2661</b>	<b>52568</b>

Table 23: Equivalence Relations ES

<i>Equivalence Relations</i>	<i>Nouns</i>	<i>Verbs</i>	<i>Total</i>
<b>EQ_NEAR_SYNONYM</b>	0	0	0
<b>EQ_SYNONYM</b>	19478	3534	23012
<b>EQ_HAS_HYPERONYM</b>	39	0	39
<b>EQ_HAS_HYPONYM</b>	14	0	14
<b>EQ_INVOLVED</b>	0	0	0
<b>EQ_IS_CAUSED_BY</b>	0	0	0
<b>EQ_HAS_HOLONYM</b>	1	0	1
<b>EQ_HAS_MERONYM</b>	2	0	2
<b>Total</b>	19534	3534	23068

The next table indicates the reliability of generated translation:

Table 24: Reliability of Equivalence Relations ES

<i>Confidence (Variants)</i>	<i>Nouns</i>	<i>Verbs</i>	<i>Total</i>
<b>100% (Manual)</b>	7819	8394	16213
<b>&gt;96%</b>	382	0	382
<b>&gt;94%</b>	2948	0	2948
<b>&gt;92%</b>	1364	0	1364
<b>&gt;85%</b>	23113	0	23113
<b>&gt;84%</b>	4156	0	4156
<b>Total</b>	39782	8394	48176

## 2.4. Quantitative conclusions

The total size of the wordnets aimed at is 50,000 word senses (synset variants) which roughly corresponds with 25-30K synsets. For each synset, at least one hyperonym relation and one equivalence relation is required, other relations are optional. The next table gives an overview of the results for all 3 wordnets:

Table 25: Subset2 Overview: NL, ES, IT

	Dutch				Italian				Spanish			
	Noun	Verb	Oth	Total	Noun	Verb	Oth	Total	Noun	Verb	Oth	Total
Synsets	24337	9125	304	33766	26466	8805	1822	37093	19663	3538	0	23201
No. of senses	40535	14278	857	55670	31318	11847	1829	44994	39782	8394	0	48176
Sens./syns.	1.67	1.56	2.82	1.65	1.2	1.34	1	1.21	2.02	2.37	0	2.08
Entries	33925	8868	785	43578	20434	6445	1829	28708	22881	3324	0	26205
Sens./entry	1.19	1.61	1.09	1.28	1.3	1.83	1	1.56	1.74	2.53	0	1.84
LIRels.	56376	22815	402	79593	66465	28463	2037	96965	43151	6756	2661	52568
LIRels/syns	2.3	2.5	1.3	2.3	2.51	3.23	1.11	2.6	2.19	1.91		2.27
EQRels-ILI	31237	14092	n.a.	45329	12176	3266	-	15442	19534	3534	0	23068
EQRels/syn	1.28	1.54	n.a.	1.34	0.5	0.37	-	0.43	0.99	1.00	0	0.99
Synsets without ILI	5022	876	n.a.	5898	14855	5539	-	20394	185	4	0	189

The first conclusion to be made is that for all 3 languages the final size has more or less been reached, although there are some differences in the figures, mainly due to the different interpretations of synonymy (strict or loose). For the Italian wordnet, the number of word senses is still 5K below the 50K limit, but the number of synsets (37K) is above the expected size (25-30K). In the case of the Dutch wordnet, a similar number of synsets (33K) correlates with a much higher number of word senses (55K), and in Spanish this is even more extreme: 23K synsets versus 48K word senses. If we compare the ratio synsets-word senses with WordNet1.5, we see that the loose definition of Spanish is very close to the ratio of WordNet1.5 and that the Dutch ratio is more strict and the Italian ratio even more strict than the Dutch. These differences in synonymy are partly compensated by the NEAR\_SYNONYM relations which are widely used for Italian, less for Dutch and never for Spanish (see Table 27 below). Nevertheless, if we compare the number of variants per synsets with the ratio in Subset1 we see that the figures have become much closer. For Dutch the ratio decreased from 1.88 to 1.65, for Spanish from 2.27 to 2.08. For Italian, on the other hand, the ratio increased from 1.09 to 1.21 (due to the measures described above).

Table 26: Overview of senses and synsets for WordNet1.5.

	Noun	Verb	Oth	Total
<b>Synsets</b>	60557	11363	19671	91591
<b>Number of senses</b>	107484	25768	34965	168217
<b>Sens./syns.</b>	1.77	2.27	1.78	1.84
<b>Entries</b>	87642	14727	24151	126520
<b>Sens./entry</b>	1.23	1.75	1.45	1.33
<b>LIRels.</b>	80735	13321	34257	128313
<b>LIRels/syns</b>	1.33	1.17	1.74	1.40

Table 27: Overview of Language Internal Relations

<i>Language Internal Relations</i>	<i>Dutch</i>		<i>Italian</i>		<i>Spanish</i>	
<b>BE_IN_STATE</b>	136	0.17%	143	0,15%	0	0.00%
<b>CAUSES</b>	995	1.25%	660	0,68%	122	0.23%
<b>HAS_HYPERONYM</b>	34459	43.29%	35522	36,63%	22881	43.53%
<b>HAS_HYPONYM</b>	34459	43.29%	35522	36,63%	22881	43.53%
<b>HAS_HOLONYM</b>	188	0.24%	266	0,27%	0	0.00%
<b>HAS_HOLO_LOCATION</b>	113	0.14%	11	0,01%	0	0.00%
<b>HAS_HOLO_MADEOF</b>	131	0.16%	162	0,17%	89	0.17%
<b>HAS_HOLO_MEMBER</b>	193	0.24%	196	0,20%	217	0.41%
<b>HAS_HOLO_PART</b>	923	1.16%	489	0,50%	1257	2.39%
<b>HAS_HOLO_PORTION</b>	55	0.07%		0,00%	0	0.00%
<b>HAS_MERONYM</b>	294	0.37%	266	0,27%	0	0.00%
<b>HAS_MERO_LOCATION</b>	113	0.14%	11	0,01%	0	0.00%
<b>HAS_MERO_MADEOF</b>	132	0.17%	162	0,17%	89	0.17%
<b>HAS_MERO_MEMBER</b>	194	0.24%	196	0,20%	217	0.41%
<b>HAS_MERO_PART</b>	926	1.16%	489	0,50%	1257	2.39%
<b>HAS_MERO_PORTION</b>	56	0.07%		0,00%	0	0.00%
<b>HAS_SUBEVENT</b>	266	0.33%	149	0,15%	0	0.00%
<b>HAS_XPOS_HYPERONYM</b>	71	0.09%		0,00%	0	0.00%
<b>HAS_XPOS_HYPONYM</b>	70	0.09%		0,00%	0	0.00%
<b>INVOLVED</b>	150	0.19%	1271	1,31%	122	0.23%
<b>INVOLVED_AGENT</b>	106	0.13%	1256	1,30%	576	1.10%
<b>INVOLVED_DIRECTION</b>	4	0.01%	19	0,02%	0	0.00%
<b>INVOLVED_INSTRUMENT</b>	414	0.52%	279	0,29%	301	0.57%
<b>INVOLVED_LOCATION</b>	63	0.08%	102	0,11%	86	0.16%
<b>INVOLVED_PATIENT</b>	673	0.85%	293	0,30%	6	0.01%
<b>INVOLVED_RESULT</b>	0	0.00%	83	0,09%	0	0.00%
<b>INVOLVED_SOURCE_DIRECTION</b>	17	0.02%	60	0,06%	0	0.00%
<b>INVOLVED_TARGET_DIRECTION</b>	23	0.03%	27	0,03%	0	0.00%
<b>IS_CAUSED_BY</b>	1031	1.30%	660	0,68%	122	0.23%
<b>IS_SUBEVENT_OF</b>	275	0.35%	149	0,15%	0	0.00%
<b>IS_MANNER_OF</b>	0	0.00%	58	0,06%	0	0.00%
<b>IN_MANNER</b>	0	0.00%	58	0,06%	0	0.00%
<b>NEAR_ANTONYM</b>	480	0.60%	57	0,06%	825	1.57%
<b>NEAR_SYNONYM</b>	252	0.32%	496	0,51%	0	0.00%
<b>ROLE</b>	151	0.19%	1282	1,32%	122	0.23%
<b>ROLE_AGENT</b>	104	0.13%	1256	1,30%	576	1.10%
<b>ROLE_DIRECTION</b>	5	0.01%	19	0,02%	0	0.00%
<b>ROLE_INSTRUMENT</b>	408	0.51%	279	0,29%	301	0.57%
<b>ROLE_LOCATION</b>	61	0.08%	102	0,11%	86	0.16%
<b>ROLE_PATIENT</b>	664	0.83%	293	0,30%	6	0.01%
<b>ROLE_RESULT</b>	0	0.00%	83	0,09%	0	0.00%
<b>ROLE_SOURCE_DIRECTION</b>	16	0.02%	60	0,06%	0	0.00%
<b>ROLE_TARGET_DIRECTION</b>	20	0.03%	27	0,03%	0	0.00%
<b>STATE_OF</b>	32	0.04%	144	0,15%	37	0.07%
<b>XPOS_NEAR_ANTONYM</b>	9	0.01%		0,00%	0	0.00%
<b>XPOS_NEAR_SYNONYM</b>	861	1.08%	14398	14,85%	392	0.75%
<b>Total</b>	79593		96965		52568	
<b>Synsets</b>	33766		37093		23201	
<b>Average per synset</b>	2.36		2.61		2.27	

The degree of language internal relations has become balanced as well across the wordnets. In the case of Subset1 the number of relations for Dutch was much higher (2.8) than in Subset2 (2.3), which is due to the fact that the work for Subset1 focussed on a richer encoding of the core wordnets, whereas Subset2 involved a large-scale extension with hyponymy relations only. In the case of Spanish and Italian we see the opposite: an increase from 1.5 to 2.6 for Italian and an increase from 2.09 to 2.27 for Spanish. The result for all the wordnets is now rather close (average of 2.4 relations per synset). If we compare this with WordNet1.5 (1.4 relations per synset) we see that EuroWordNet is richer in terms of relations. The next column gives the distribution of these Language Internal Relations. The first column gives the absolute number of relations per type, the second column for each language gives the relative percentage. In all 3 languages, hyponymy is the most important relation (82-87%). The distribution of the other relations is also rather balanced, except for the XPOS\_NEAR\_SYNONYM relations for Italian (14,85%). This also explains the relatively high average of relations for Italian. For the rest there are only minor differences due to local approaches, but no striking differences. For comparison, the next table gives the distribution for WordNet1.5.

Table 28: Distribution of different relations in WordNet1.5

Relation	Nouns		Verbs		Other		Total	
<b>antonym</b>	1713	2.09%	1025	7.69%	4452	13.00%	7190.098	5.55%
<b>hyponym</b>	61123	74.47%	10817	81.20%	0	0.00%	71941.56	55.48%
<b>mero-member</b>	11472	13.98%	0	0.00%	0	0.00%	11472.14	8.85%
<b>mero-sub</b>	366	0.45%	0	0.00%	0	0.00%	366.0045	0.28%
<b>mero-part</b>	5695	6.94%	0	0.00%	0	0.00%	5695.069	4.39%
<b>entailment</b>	0	0.00%	435	3.27%	0	0.00%	435.0327	0.34%
<b>cause</b>	0	0.00%	204	1.53%	0	0.00%	204.0153	0.16%
<b>also-see</b>	0	0.00%	840	6.31%	2686	7.84%	3526.063	2.72%
<b>attribute</b>	1713	2.09%	0	0.00%	636	1.86%	2349.021	1.81%
<b>Similar to</b>	0	0.00%	0	0.00%	20050	58.53%	20050	15.46%
<b>Pertains to</b>	0	0.00%	0	0.00%	6433	18.78%	6433	4.96%
<b>Total</b>	82082		13321		34257		129660	

Finally, Table29 lists the equivalence relations for Dutch, Italian and Spanish.

Table 29: Overview of Equivalence Relations

Equivalence Relations	Dutch			Spanish			Italian		
	Nouns	Verbs	Total	Nouns	Verbs	Total	Nouns	Verbs	Total
EQ_SYNONYM	1730	275	2005	19478	3534	23012	7225	773	7998
EQ_NEAR_SYNONYM	28816	13190	42006	0	0	0	1595	1898	2493
EQ_HAS_HYPERONYM	446	564	1010	39	0	39	3314	583	3897
EQ_HAS_HYPONYM	140	20	160	14	0	14	13	12	25
EQ_INVOLVED	2	13	150	0	0	0	7		7
EQ_ROLE	9	0	9	0	0	0	0	0	0
EQ_IS_CAUSED_BY	3	15	18	0	0	0	11		11
EQ_CAUSES	8	8	16	0	0	0	0	0	0
EQ_HAS_HOLONYM	48	0	48	1	0	1	2		2
EQ_HAS_MERONYM	21	0	21	2	0	2	8		8
EQ_HAS_SUBEVENT	0	2	2	0	0	0	0	0	0
EQ_IS_SUBEVENT_OF	0	3	3	0	0	0	0	0	0
EQ_BE_IN_STATE	0	0	0	0	0	0	1		1
<b>Total</b>	<b>31237</b>	<b>14092</b>	<b>45329</b>	<b>19534</b>	<b>3534</b>	<b>23068</b>	<b>12176</b>	<b>3266</b>	<b>15442</b>

The main conclusions we can make here are:

- no usage of EQ\_NEAR\_SYNONYM in the Spanish wordnet
- extreme usage of EQ\_NEAR\_SYNONYM in the Dutch wordnet
- extreme usage of EQ\_HAS\_HYPERONYM in Italian
- low number of equivalence relations for Italian

The difference in number for the EQ\_NEAR\_SYNONYM relations mainly have to do with the choice for Dutch to assign this relation type to all the automatically derived equivalences. For the Spanish wordnet, automatically extracted equivalences are of the type EQ\_SYNONYM. In the Italian wordnet, automatically generated equivalences have to be verified manually before they can be assigned to synsets, which explains the lower figure of equivalence links. If not (yet) linked to the ILI, these Italian synsets are linked with EQ\_HAS\_HYPERONYM links to their superordinate translation. The same holds for many Italian entries are not found within the bilingual Italian-English dictionary. This explains the large number of these equivalence relations in the Italian wordnet.

### 3. Coverage of Subet2 over top concept clusters

As explained in D014D015 (Vossen et al. 1998), the wordnets are built top-down starting with the Base Concepts. Each site is free to include different lexicalizations patterns when extending the vocabulary from the Base Concepts down. Still, to get an idea of the conceptual distribution of this extension we also measure the progress of the wordnets relative to the Top Ontology, which represents the diversity of Base Concepts that have been selected. For this purpose, AMS implemented an inheritance mechanism that derives the Top Concepts from hyperonyms in WordNet1.5. By loading ILI-equivalences of the Spanish, Dutch and Italian first subset in the Amsterdam lexical database (ALS), it is possible to collect the Top Concepts that apply to these equivalences via hyponymy-inheritance in WordNet1.5. By applying this to all the equivalences, it is possible to quantify the coverage per top concept. Note that this measurement depends on the quality and quantity of the equivalence relations. Not all synsets in the Italian and Dutch wordnets have a (correct) equivalent relation. Furthermore, it may be that the hyponymy relations in the local wordnets are different, but the global semantic classification still has to be consistent. This method therefore still gives a good indication of the conceptual coverage.

The Top Ontology is divided in 3 main parts:

- 1stOrderEntities (nouns): concrete things
- 2ndOrderEntities (nouns, verbs and adjectives): states, events, processes, relations and properties
- 3rdOrderEntities (nouns): idea, knowledge, propositions

The results are given in the next tables, where nouns are divided into separate tables for 1st, 2nd and 3rdOrder Entities, and the verbs listed in one 2ndOrderEntity table. It should be noted that what we quantify is not the number of synsets but the number of Top-Concept assignment. Due to inheritance and multiple Top-Concept assignments, most synsets get several Top-Concepts. A Top-Concept is however only assigned once if it is derived via multiple paths or nodes.

In Table 30, the results are given for the 1st Order Entities. The first column lists the 1stOrder Top-Concepts. The next column gives the number of synset-assignments in WordNet1.5: are either directly or indirectly (via a hyperonym chain). The 3rd column gives the percentages of the total clusters in WordNet1.5. The 1st column of each wordnet gives the TC-clustering for Subset1 (S1) based on the TC-inheritance in WordNet1.5 of the ILI-records representing the local wordnet synsets. The 2nd column of each wordnet gives the clustering for Subset 2 (S2). The difference

between S1 and S2 indicates the progress. The next column gives the percentage of the total set of 1stOrder nouns covered by each wordnet and the 4th column for NL, ES and IT gives the percentage of the corresponding TC cluster in WordNet1.5.

Table 30: Nominal Synsets clustered as 1stOrder Concepts

Top-Concept	WN		AMS				FUE				PSA			
	Syns	% of wn	S1	S2	% of nl-net	% of wn-cluster	S1	S2	% of es-net	% of wn-cluster	S1	S2	% of it-net	% of wn-cluster
Animal	6958	3,53%	65	469	1,10%	6,7%	782	833	1,46%	12,0%	310	321	1,89%	4,6%
Artifact	12446	6,31%	1234	3852	8,99%	30,9%	4496	4630	8,13%	37,2%	626	1227	7,21%	9,9%
Building	589	0,30%	105	264	0,62%	44,8%	282	282	0,50%	47,9%	5	65	0,38%	11,0%
Comestible	2304	1,17%	161	522	1,22%	22,7%	579	938	1,65%	40,7%	77	131	0,77%	5,7%
Container	1060	0,54%	59	251	0,59%	23,7%	321	317	0,56%	29,9%	36	50	0,29%	4,7%
Covering	1279	0,65%	103	505	1,18%	39,5%	520	529	0,93%	41,4%	10	138	0,81%	10,8%
Creature	473	0,24%	2	81	0,19%	17,1%	50	51	0,09%	10,8%	2	4	0,02%	0,8%
Function	21284	10,79%	1569	6208	14,50%	29,2%	7496	8157	14,32%	38,3%	1265	2534	14,90%	11,9%
Furniture	196	0,10%	17	54	0,13%	27,6%	68	74	0,13%	37,8%	4	32	0,19%	16,3%
Garment	446	0,23%	22	226	0,53%	50,7%	195	202	0,35%	45,3%	2	74	0,43%	16,6%
Gas	56	0,03%	8	37	0,09%	66,1%	26	28	0,05%	50,0%	11	12	0,07%	21,4%
Group	13113	6,65%	226	976	2,28%	7,4%	1432	1558	2,74%	11,9%	332	461	2,71%	3,5%
Human	6862	3,48%	220	2085	4,87%	30,4%	2827	2892	5,08%	42,1%	622	1282	7,54%	18,7%
ImageRep	480	0,24%	28	135	0,32%	28,1%	171	173	0,30%	36,0%	3	11	0,06%	2,3%
Instrument	4557	2,31%	512	1417	3,31%	31,1%	1676	1696	2,98%	37,2%	480	688	4,04%	15,1%
Language Rep	1883	0,95%	107	475	1,11%	25,2%	527	535	0,94%	28,4%	15	99	0,58%	5,3%
Liquid	1083	0,55%	67	210	0,49%	19,4%	229	240	0,42%	22,2%	45	48	0,28%	4,4%
Living	23704	12,02%	610	3357	7,84%	14,2%	4801	5011	8,80%	21,1%	1215	1947	11,44%	8,2%
MoneyRep	241	0,12%	23	68	0,16%	28,2%	81	75	0,13%	31,1%	4	8	0,05%	3,3%
Natural	37427	18,98%	2124	7673	17,92%	20,5%	9730	10211	17,93%	27,3%	1914	3264	19,19%	8,7%
Object	27603	14,00%	2046	7125	16,64%	25,8%	8987	9443	16,58%	34,2%	1729	3045	17,90%	11,0%
Occupation	1222	0,62%	42	357	0,83%	29,2%	571	581	1,02%	47,5%	164	264	1,55%	21,6%
Part	7412	3,76%	587	1762	4,11%	23,8%	1957	2282	4,01%	30,8%	72	198	1,16%	2,7%
Place	3253	1,65%	223	759	1,77%	23,3%	858	1049	1,84%	32,2%	39	103	0,61%	3,2%
Plant	8221	4,17%	90	363	0,85%	4,4%	715	798	1,40%	9,7%	262	310	1,82%	3,8%
Representat	592	0,30%	55	198	0,46%	33,4%	245	251	0,44%	42,4%	10	35	0,21%	5,9%
Software	134	0,07%	5	28	0,07%	20,9%	29	29	0,05%	21,6%	2	4	0,02%	3,0%
Solid	3985	2,02%	324	1121	2,62%	28,1%	1157	1257	2,21%	31,5%	38	180	1,06%	4,5%
Substance	7912	4,01%	704	2097	4,90%	26,5%	2263	2640	4,64%	33,4%	154	373	2,19%	4,7%
Vehicle	453	0,23%	38	149	0,35%	32,9%	189	186	0,33%	41,1%	66	104	0,61%	23,0%
Total	197228		11376	42824		21,7%	53260	56948		28,9%	9514	17012		8,6%

If the wordnets are equally balanced then the relative percentages of the wordnets should be the same, even if the total size of the wordnets are different. When a particular percentage is significantly lower than the other wordnets it means that this wordnet should be extended in this domain to become more balanced.<sup>5</sup> If WordNet1.5 is used as a comparison, the percentage of the 4th column should be about 30%, since the total size of the wordnets is about 1/3 of WordNet1.5. However, as mentioned before, some areas such as Animal and Plant are very difficult to match because WordNet1.5 contains a lot of expert terminology in these particular domains. Furthermore, we should realize that these clusterings are based on the ILI-equivalences linked to the synsets. If

<sup>5</sup> The table is also useful for users of the wordnets to verify if particular domains or fields of their interest are well-represented or need to be extended.

no equivalences are given, we cannot derive Top-Concept assignments for this synset.

First of all we see, as expected, that Creature, Animal, and Plant are less well covered in all 3 wordnets, if compared to WordNet1.5. Another, unexpected, case of under-specification is Group. This is an issue which we have to look into. For Spanish and Dutch, all other clusters are well-represented. In general we can say that the Dutch and Spanish wordnets are well-balanced with respect to WordNet1.5 and also with respect to each other. Some classes are even over-represented in both wordnets: Function, Garment, Gas, Natural and Substance. The Italian clustering is rather poor, but this is due to the small amount of equivalence relations. The equivalence relations have to be completed before we can make any conclusions for Italian.

The next two tables shows the distribution for nouns and verbs that are classified as 2ndOrderEntities according to the WordNet1.5 hyponymy chains.

Table 31: Nominal Synsets clustered as 2ndOrder Concepts

Top-Concept	WN		AMS				FUE				PSA			
	Syns	% of wn	S1	S2	% of nl-net	% of wn-cluster	S1	S2	% of es-net	% of wn-cluster	S1	S2	% of it-net	% of wn-cluster
Agentive	7214	6,80%	404	2159	7,31%	29,9%	3039	3139	7,22%	43,5%	91	942	10,64%	13,1%
BoundedEvent	4753	4,48%	292	1451	4,91%	30,5%	1934	2015	4,63%	42,4%	58	551	6,22%	11,6%
Cause	9071	8,56%	595	2727	9,23%	30,1%	3862	3970	9,13%	43,8%	137	1069	12,07%	11,8%
Communication	4325	4,08%	256	1235	4,18%	28,6%	1591	1618	3,72%	37,4%	45	308	3,48%	7,1%
Condition	2325	2,19%	311	701	2,37%	30,2%	994	1009	2,32%	43,4%	28	176	1,99%	7,6%
Dynamic	11760	11,09%	741	3459	11,71%	29,4%	4989	5167	11,88%	43,9%	193	1280	14,45%	10,9%
Existence	198	0,19%	19	76	0,26%	38,4%	97	98	0,23%	49,5%	1	29	0,33%	14,6%
Experience	4012	3,78%	294	1107	3,75%	27,6%	1754	1795	4,13%	44,7%	88	344	3,88%	8,6%
Location	851	0,96%	60	259	0,88%	30,4%	343	346	0,80%	40,7%	11	109	1,23%	12,8%
Manner	573	0,54%	29	162	0,55%	28,3%	241	247	0,57%	43,1%	3	51	0,58%	8,9%
Mental	6275	5,92%	396	1734	5,87%	27,6%	2482	2534	5,82%	40,4%	94	484	5,47%	7,7%
Modal	291	0,27%	17	88	0,30%	30,2%	130	132	0,30%	45,4%	5	14	0,16%	4,8%
Phenomenal	1227	1,16%	144	331	1,12%	27,0%	508	515	1,18%	42,0%	32	92	1,04%	7,5%
Physical	4714	4,45%	445	1369	4,64%	29,0%	1915	1938	4,45%	41,1%	80	317	3,58%	6,7%
Possession	889	0,84%	64	225	0,76%	25,3%	275	277	0,64%	31,2%	9	43	0,49%	4,8%
Property	6999	6,60%	508	1827	6,19%	26,1%	3173	3233	7,43%	46,2%	101	393	4,44%	5,6%
Purpose	9250	8,73%	459	2545	8,62%	27,5%	3374	3433	7,89%	37,1%	95	819	9,25%	8,9%
Quantity	2228	2,10%	99	494	1,67%	22,2%	686	720	1,65%	32,3%	24	71	0,80%	3,2%
Relation	4154	3,92%	248	1083	3,67%	26,1%	1472	1498	3,44%	36,1%	48	253	2,86%	6,1%
Social	7449	7,03%	353	1972	6,68%	26,5%	2650	2695	6,19%	36,2%	71	519	5,86%	7,0%
Static	12522	11,81%	850	3198	10,83%	25,5%	5069	5150	11,84%	41,1%	178	701	7,92%	5,6%
Stimulating	599	0,57%	42	199	0,67%	33,2%	298	296	0,68%	49,4%	7	22	0,25%	3,7%
Time	822	0,78%	34	200	0,68%	24,3%	265	300	0,69%	36,5%	10	24	0,27%	2,9%
UnboundedEvent	2792	2,63%	124	818	2,77%	29,3%	1226	1244	2,86%	44,6%	39	242	2,73%	8,7%
Usage	723	0,68%	14	114	0,39%	15,8%	138	136	0,31%	18,8%	0	3	0,03%	0,4%
Total	106016		6798	29533		27,9%	42505	43505		41,0%	1448	8856		8,4%

Table 32: Verbal Synsets clustered as 2ndOrder Concepts

Top-Concept	WN		AMS				FUE				PSA			
	Syns	% of wn	S1	S2	% of NL-net	% of wn-cluster	S1	S2	% of ES-net	% of wn-cluster	S1	S2	% of IT-net	% of wn-cluster
Agentive	3520	7,0%	377	1418	7,3%	40,3%	815	923	6,4%	26,2%	77	353	7,2%	10,0%
BoundedEvent	4563	9,1%	536	1871	9,6%	41,0%	1209	1326	9,2%	29,1%	99	431	8,8%	9,4%
Cause	6802	13,5%	799	2742	14,1%	40,3%	1801	1981	13,8%	29,1%	211	714	14,5%	10,5%
Communication	1618	3,2%	159	656	3,4%	40,5%	399	464	3,2%	28,7%	46	189	3,8%	11,7%
Condition	727	1,4%	79	343	1,8%	47,2%	208	221	1,5%	30,4%	17	72	1,5%	9,9%
Dynamic	10312	20,5%	1363	4114	21,2%	39,9%	2813	3101	21,6%	30,1%	342	1072	21,8%	10,4%
Existence	1028	2,0%	103	405	2,1%	39,4%	261	270	1,9%	26,3%	21	108	2,2%	10,5%
Experience	789	1,6%	91	317	1,6%	40,2%	303	312	2,2%	39,5%	32	100	2,0%	12,7%
Location	3711	7,4%	522	1528	7,9%	41,2%	1190	1310	9,1%	35,3%	111	338	6,9%	9,1%
Manner	174	0,3%	21	67	0,3%	38,5%	52	52	0,4%	29,9%	4	10	0,2%	5,7%
Mental	1147	2,3%	107	432	2,2%	37,7%	309	335	2,3%	29,2%	31	139	2,8%	12,1%
Modal	36	0,1%	6	15	0,1%	41,7%	11	15	0,1%	41,7%	0	3	0,1%	8,3%
Phenomenal	55	0,1%	7	28	0,1%	50,9%	24	27	0,2%	49,1%	0	3	0,1%	5,5%
Physical	5247	10,4%	692	2157	11,1%	41,1%	1653	1765	12,3%	33,6%	177	502	10,2%	9,6%
Possession	879	1,7%	106	336	1,7%	38,2%	205	247	1,7%	28,1%	18	84	1,7%	9,6%
Property	194	0,4%	17	74	0,4%	38,1%	47	50	0,3%	25,8%	2	16	0,3%	8,2%
Purpose	1958	3,9%	191	782	4,0%	39,9%	397	447	3,1%	22,8%	34	192	3,9%	9,8%
Quantity	303	0,6%	40	117	0,6%	38,6%	93	102	0,7%	33,7%	6	23	0,5%	7,6%
Relation	388	0,8%	44	146	0,8%	37,6%	100	113	0,8%	29,1%	9	39	0,8%	10,1%
Social	2362	4,7%	260	874	4,5%	37,0%	458	531	3,7%	22,5%	56	239	4,9%	10,1%
Static	2720	5,4%	301	336	1,7%	12,4%	803	258	1,8%	9,5%	46	87	1,8%	3,2%
Stimulating	367	0,7%	29	140	0,7%	38,1%	169	174	1,2%	47,4%	28	49	1,0%	13,4%
Time	42	0,1%	1	10	0,1%	23,8%	6	7	0,0%	16,7%	0	1	0,0%	2,4%
UnboundedEvent	1055	2,1%	146	454	2,3%	43,0%	260	291	2,0%	27,6%	29	137	2,8%	13,0%
Usage	244	0,5%	36	68	0,3%	27,9%	58	40	0,3%	16,4%	3	14	0,3%	5,7%
Total	50241		6033	19430		38,7%	13644	14362		28,6%	1399	4915		9,8%

For the 2ndOrderEntities the situation is even better. For most clusters the Dutch and Spanish wordnets score around 30% of the WordNet1.5 coverage. The Dutch verbs even score a bit higher on average. We do find a negative score for all wordnets for Usage (2ndOrder nouns) and Static (verbs). We also see that the proportion is less than in WordNet1.5 (Usage nouns: 0.3% in EWN and 0.68% in WordNet; Static verbs: 1,8% in EWN versus 5.4% in WordNet1.5). These are the only areas that need more balancing.

Finally, the next table gives the nominal synsets classified as 3rdOrderEntities, where the percentage give the proportion of the set in WordNet1.5.

Table 33: Nominal Synsets clustered as 3rdOrder Concepts

	WN	AMS			FUE			PSA		
	Syns	S1	S2	% of wn	S1	S2	% of wn	S1	S2	% of wn
<b>3rdOrderEntity</b>	4989	309	1388	27.82%	1860	1912	38.32%	147	340	6.81%

Here we see the same tendency as above: the Spanish and Dutch wordnet neatly cover 1/3 of WordNet1.5 and that the Italian wordnet lacks equivalence relations.

Since we also added the WordNet1.5 lexicographer's file codes to the database it is also possible to measure the subsets with respect to that classification. This is shown in the next table:

Table 34: Subset2 clustered over the WordNet1.5 Lexicographer's file codes

WN Lexicographer's file code	WN	% of WN-net	AMS	% of NL-net	FUE	% of ES-net	PSA	% of IT-net	
4	noun.act	4953	5,92%	1475	7.38%	2143	8.68%	729	11.58%
5	noun.animal	6742	8,06%	378	1.89%	718	2.91%	313	4.97%
6	noun.artifact	9420	11,26%	3255	16.28%	3730	15.11%	1090	17.31%
7	noun.attribute	2553	3,05%	573	2.87%	1256	5.09%	82	1.30%
8	noun.body	1692	2,02%	445	2.23%	461	1.87%	29	0.46%
9	noun.cognition	2409	2,88%	705	3.53%	1077	4.36%	171	2.72%
10	noun.communication	4144	4,95%	1181	5.91%	1556	6.30%	289	4.59%
11	noun.event	787	0,94%	222	1.11%	376	1.52%	53	0.84%
12	noun.feeling	361	0,43%	85	0.43%	192	0.78%	62	0.98%
13	noun.food	2293	2,74%	511	2.56%	927	3.76%	125	1.99%
14	noun.group	7273	8,69%	499	2.50%	600	2.43%	57	0.91%
15	noun.location	1889	2,26%	350	1.75%	556	2.25%	54	0.86%
16	noun.motive	24	0,03%	7	0.04%	15	0.06%		0.00%
17	noun.object	2472	2,96%	587	2.94%	659	2.67%	46	0.73%
18	noun.person	5371	6,42%	1623	8.12%	2383	9.65%	1169	18.57%
19	noun.phenomenon	444	0,53%	128	0.64%	154	0.62%	31	0.49%
20	noun.plant	7933	9,48%	273	1.37%	667	2.70%	300	4.76%
21	noun.possession	794	0,95%	180	0.90%	232	0.94%	25	0.40%
22	noun.process	579	0,69%	121	0.61%	252	1.02%	51	0.81%
23	noun.quantity	1159	1,39%	202	1.01%	287	1.16%	11	0.17%
24	noun.relation	525	0,63%	129	0.65%	199	0.81%	14	0.22%
25	noun.shape	383	0,46%	99	0.50%	150	0.61%	11	0.17%
26	noun.state	1857	2,22%	482	2.41%	753	3.05%	68	1.08%
27	noun.substance	2624	3,14%	736	3.68%	808	3.27%	88	1.40%
28	noun.time	810	0,97%	197	0.99%	293	1.19%	24	0.38%
29	verb.body	420	0,50%	182	0.91%	150	0.61%	55	0.87%
30	verb.change	2827	3,38%	1137	5.69%	880	3.56%	317	5.03%
31	verb.cognition	771	0,92%	298	1.49%	168	0.68%	82	1.30%
32	verb.communication	1445	1,73%	597	2.99%	430	1.74%	168	2.67%
33	verb.competition	366	0,44%	105	0.53%	41	0.17%	20	0.32%
34	verb.consumption	220	0,26%	95	0.48%	66	0.27%	17	0.27%
35	verb.contact	1859	2,22%	698	3.49%	634	2.57%	142	2.26%
36	verb.creation	757	0,90%	291	1.46%	200	0.81%	65	1.03%
37	verb.emotion	268	0,32%	96	0.48%	139	0.56%	48	0.76%
38	verb.motion	1750	2,09%	766	3.83%	595	2.41%	157	2.49%
39	verb.perception	378	0,45%	149	0.75%	141	0.57%	35	0.56%
40	verb.possession	814	0,97%	313	1.57%	224	0.91%	74	1.18%
41	verb.social	1700	2,03%	651	3.26%	407	1.65%	181	2.87%
42	verb.stative	532	0,64%	158	0.79%	145	0.59%	41	0.65%
43	verb.weather	50	0,06%	18	0.09%	23	0.09%	2	0.03%
<b>Total</b>		83648		19997		24687		6296	

This table more or less confirms what has been concluded above. The distribution of Spanish and Dutch is compatible with Wordnet1.5 except for Plant, Animal and Group. Italian figures cannot be interpreted because of the lack of equivalences.

## 4. Comparison of Subset2

### 4.1. Introduction

There are two main objectives of the comparison:

- 1) to measure the degree of coverage and intersection of subset2 in the same way this was carried out for subset1
- 2) to evaluate the improvement (if any) of the intersection figures compared with the figures obtained for subset1.

For this comparison each site (NL, IT, SP) has generated two sets (one for nouns and one for verbs) of hyponymy chains. For example, the next list of Dutch senses is generated for "opstijgen" (take off) by recursively taking all the hyperonyms:

- opstijgen (take off) stijgen (move to a higher position) verplaatsen (move location)  
voortbewegen (move location) bewegen (move reflexive) bewegen (move intransitive)  
veranderen (change)

To be able to compare these chains, each word sense in the chain has been replaced by the ILI-records that are linked to these synsets by eq\_synonym and eq\_near\_synonym relations. When we reverse this chain (from top to bottom) we get the following result:

00064108 01046072 01046072 01046072 01055491 01094615 00257753

This means that the Dutch equivalent to ILI record number 00064108 has as hyponym (the equivalent to ILI record number 01046072) and this one has as hyponym (the equivalent to ILI record number 01046072), etc.<sup>6</sup> Two kinds of measurements have been applied to these chains:

- Edge-coverage of chains means that not only the synsets but also the hyponymy relations between them are covered by the different wordnets.
- Node-coverage of chains means that the synsets are covered but not necessarily the hyponymy relations. Perhaps another relation holds between the corresponding synsets or perhaps they are unrelated.

Both measurements are important and can be used in different way. Of course edge-coverage is more difficult to achieve (covering an edge implies covering the two related nodes and the relation between them -in the same direction-). A high degree of edge-covering overlap means that the overlapping concepts exist and are lexicalized in all the languages that overlap and that their structural (hyponym/hyperonym) relationships hold in the same way for such languages. A lower level of edge-covering overlapping could indicate:

- a) incompleteness in covering the nodes (can be measured by node-coverage)
- b) incompleteness of relations in the language (can be measured by edge-coverage)
- c) A genuine difference between languages

---

<sup>6</sup> In some cases (all of them for Dutch and Italian) an ILI chain contains nodes that have not been linked to WN1.5 equivalents. In these cases the original word instead of the ILI record number was used to identify the node.

Complete overlapping of chains (either at edge or node level) is difficult to achieve in the case of huge differences in the size of the wordnets to be compared (e.g. the nouns in the Spanish wordnet, which is the largest subset, still hardly covers 30% of the nouns in WN1.5). As stated above, complete compatibility may not be the goal, eventually. There are differences at the highest level of the hierarchy that are based on different insights.

We have therefore developed two other kind of measurements that are more useful for comparison: subsequences and subsequences with gaps:

- Subsequences are simply chains of nodes/edges that exactly match a fragment of another chain. Subsequences can be classified according to their length.
- Subsequences with  $n$  gaps are chains of nodes/edges that match a fragment of another chain but failing to match  $n$  nodes of edges.

Further details on the comparison can be found in D014D015 (Vossen et al. 1998). As in the case of subset1 the statistics have been extracted at three levels:

- 1) Individual level (data provided by each site without any cross comparison).
- 2) Degree of coverage of WN1.5.
- 3) Overlapping with the other sites.

Because of the fine-grained sense-distinction in WordNet1.5 it is possible that differences in the ILI-chains are the result of choosing different but related senses. Furthermore, some of the sites have assigned multiple translations to these related senses but had to choose one for generating the ILI-chains. If all combinations of chains were generated the number of chains would be too high. We have therefore applied a new experiment using composite ILIs (of the type generalisation) instead of the original ILI records. The aim of this experiment is to check to what extent clustering of obviously related synsets (i.e. a composite ILI) would increase the overlapping between languages.

## 4.2. Evaluation of individual wordnets

Table 35 ILI chains for nouns

	<i>ILI nodes</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>CHAINS</i>
<b>ES</b>	19800	11	15056	4734	20124	17768
<b>IT</b>	4111	24	3859	257	6484	11546
<b>NL</b>	20050	237	16024	4620	23437	37753
<b>WN15</b>	60557	11	47110	13436	61123	53467

Table 36 ILI chains for verbs

	<i>ILI nodes</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>CHAINS</i>
<b>ES</b>	3507	355	2548	784	3166	2565
<b>IT</b>	1391	57	1193	170	2206	2443
<b>NL</b>	6382	2	4951	1867	8755	9564
<b>WN15</b>	11363	573	8446	2580	10816	8486

Table 37 ILI chains (total)

	<i>ILI nodes</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>CHAINS</i>
<b>ES</b>	23307	366	17604	5518	23290	20333
<b>IT</b>	5502	81	5052	427	8690	13989
<b>NL</b>	26432	239	20975	6487	32222	47317
<b>WN15</b>	71920	584	55556	16016	71939	61953

What must be pointed out here is the increment of number of nodes with regard to subset1, from 21,795 to 23,307 (7%) for Spanish, from 2,149 to 5,502 (156%) for Italian and from 7,240 to 26,432 (265%) for Dutch. The lower increment for Spanish is of course due to the fact that the objectives of the projects regarding the size of Spanish WN were almost reached with subset1 and so, as has been pointed out in section 2.3, improving the quality and not the size has been the main objective of the extension. With regard to the relatively lower figures for Italian two factors must be pointed out:

1. The high (in absolute and relative terms) number of Italian synsets without ILI equivalent (20,193 from 35,273, i.e. 57%, when the figures for Dutch are 5,898 from 33,766, i.e. 17%, and for Spanish, 189 from 23,201, i.e. 1%). As the comparison experiments measure the degree of coverage of WN1.5 chains all the synsets (and the chains where those synsets occur) not related to ILI (WN1.5) by means of equivalent relation are not considered for comparison.
2. The high number (in absolute and relative terms) of isolated synsets. The total numbers of Italian synsets related to ILI are 10,463 for nouns and 2,800 for verbs. But from these only 4,111 noun synsets (39%) and 1,391 verb synsets (49%) occur in some chain, the rest being isolated nodes. As the comparison experiments measure coverage of chains (length 0 chains are not considered) a great amount of Italian synsets do not occur in tables 34, 35 and 36.

The next two tables present the number and % of noun and verb chains classified by length for each language.

Table 38 Frequencies and ratios of noun chains / length / language

	WN		NL		IT		ES	
	frequency	%	frequency	%	frequency	%	frequency	%
<b>1</b>			139	0.15	2	0.02		
<b>2</b>	33	0.06	52	0.06	1439	12.92	101	0.57
<b>3</b>	522	0.97	953	1.04	1876	16.84	761	4.29
<b>4</b>	2231	4.15	5407	5.89	1284	11.53	1791	10.09
<b>5</b>	5695	10.58	13356	14.54	2743	24.63	2929	16.50
<b>6</b>	12781	23.75	14577	15.87	2063	18.52	4057	22.86
<b>7</b>	11804	21.94	15089	16.43	349	3.13	4007	22.58
<b>8</b>	8787	16.33	13044	14.20	822	7.38	2375	13.38
<b>9</b>	6005	11.16	11139	12.13	544	4.88	1064	6.00
<b>10</b>	3358	6.24	8469	9.22	16	0.14	462	2.60
<b>11</b>	1415	2.63	5758	6.27			128	0.72
<b>12</b>	519	0.96	2607	2.84			45	0.25
<b>13</b>	367	0.68	852	0.93			23	0.13
<b>14</b>	214	0.04	295	0.32			2	0.01
<b>15</b>	75	0.14	90	0.10			2	0.01
<b>16</b>	7	0.01	24	0.03				
<b>17</b>			4	0.00				
<b>Total</b>	53813	100	91855	100	11138	100	17747	100
<b>Average</b>	7.19		7.46		4.83		6.32	

Table 39 Frequencies and ratios of verb chains / length /language

	WN		NL		IT		ES	
	frequency	%	frequency	%	frequency	%	frequency	%
1	236	2.78			1	0.03	180	7.02
2	1867	22.00	40	0.07	841	26.15	740	28.85
3	2532	29.83	382	0.66	710	22.08	755	29.43
4	1959	23.08	1089	1.88	487	15.14	501	19.53
5	1028	12.11	3161	5.47	418	13.00	233	9.08
6	463	5.45	5571	9.63	372	11.57	94	3.66
7	250	2.95	8074	13.96	265	8.24	43	1.68
8	109	1.28	8334	14.41	102	3.17	18	0.70
9	32	0.38	7558	13.07	20	0.62	1	0.04
10	10	0.12	6308	10.91				
11	2	0.02	5596	9.68				
12			4378	7.57				
13			3001	5.19				
14			1586	2.74				
15			908	1.57				
16			543	0.94				
17			310	0.54				
18			241	0.42				
19			264	0.46				
20			151	0.26				
21			158	0.27				
22			72	0.12				
23			60	0.10				
24			33	0.06				
25			9	0.02				
26			9	0.02				
<b>Total</b>	8488	100	57836	100	3216	100	2565	100
<b>Average</b>	3.58		9.20		4.02		3.16	

There has been an improvement on the connectivity of the individual WNs measured by the average length of the chains. For Spanish the average length of the noun chains increases from 6.22 to 6.32 (2%) and in the case of verbs from 3.01 to 3.16 (5%). For Italian the average length of noun chains from 4.15 to 4.83 (16%) and for verbs from 3.69 to 4.02 (9%). For Dutch the average length of noun chains increases from 6.22 to 7.46 and for verbs decreases from 9.59 down to 9.20.

There has been a clear improvement of Italian figures and the results are now closer to the WN1.5 distribution. The reasons that we pointed out in D014D015 (Vossen et al. 1998) for explaining the low average length of Italian figures, seem to be correct. The lower increment of average length for Spanish are due to 1) the low increment in size of Spanish WN and 2) the fact that average lengths of the Spanish WN are already quite closed to WN1.5 (6.32 against 7.19 for nouns and 3.16 against 3.58 for verbs).

In the case of Dutch, the results are more difficult to be evaluated. For nouns the current average length of chains just exceeds the WN1.5 result (7.46 vs 7.19). In the case of verbs the differences are greater (9.20 vs 3.58). This problem was addressed when assessing subset1 results and we detected that in some cases this corresponds to pathological chains due to incorrect equivalence relations. As explained in section 2.1, we decided to manually analyse the most frequent long chains for correctness of the equivalence relations. We can now see that the situations has improved with respect to Subset1: the results are now better despite the fact that the number of ILI nodes has increased 3 times in size). For instance, the average length of chains for verbs has decreased from

9.59 down to 9.20 and the length of chains overcoming 1% of frequency has decreased from 17 to 15. However, we can also conclude that an additional effort has to be done to further improve the translations.

### 4.3. Global evaluation

The next three tables account for the coverage of the individual wordnets (NL, IT, SP), pairs (NL-IT, NL-SP, IT-SP) and full intersection (NL-IT-SP) against 1) WN1.5 and 2) the union NL-IT-SP. The results have to be compared with the corresponding (tables 37, 38 and 39 in D014D015, Vossen et al. 98).<sup>7</sup> The results are clearly better both in absolute and relative terms. For two languages (Dutch and Spanish) the % of coverage with WN1.5 is around 30%. The coverage of intersection of Dutch and Spanish is also remarkable (more than 13% vs. 4.81% for subset1). The absolute figures of Italian coverage are lower because of the small amount of equivalence relations for the Italian WN (as discussed in section 4.2). However, the figures for the intersection of the three languages (Dutch, Italian and Spanish) have raised from 1.89% to 13.21%. These figures have to be considered as good, taking into account that improving the intersection was not the main objective of subset2 extension.

Table 40 Coverage of noun ILI records

	Total	(62780)	(26392)
	frequency	% $\cup$ (WN,IT, NL, ES)	% $\cup$ (IT, NL, ES)
<b>ES</b>	19663	31.32%	74.50%
<b>IT</b>	6338	10.10%	24.01%
<b>NL</b>	16107	25.66%	61.03%
$\cap$ (ES, IT)	4719	7.52%	17.88%
$\cap$ (ES, NL)	10123	16.12%	38.36%
$\cap$ (IT, NL)	4142	6.60%	15.69%
$\cap$ (ES, IT, NL)	3318	5.29%	12.57%

Table 41 Coverage of verb ILI records

	Total	(12215)	(6856)
	frequency	% $\cup$ (WN,IT, NL, ES)	% $\cup$ (IT, NL, ES)
<b>ES</b>	3538	28.96%	51.60%
<b>IT</b>	2332	19.09%	34.01%
<b>NL</b>	5358	43.86%	78.15%
$\cap$ (ES, IT)	1288	10.54%	18.79%
$\cap$ (ES, NL)	2438	19.96%	35.56%
$\cap$ (IT, NL)	1721	14.09%	25.10%
$\cap$ (ES, IT, NL)	1075	8.80%	15.68%

<sup>7</sup> When comparing the results with the tables in D014D015 for subset1, it should be noted that there has been a slight change in the procedure. The ILI-chains do not exactly represent the complete set of ILI-records related to the wordnets. Because of technical complications, only one translation is chosen when multiple translations are linked to a synset. In the case of the Spanish wordnet this will not make much difference since most synsets only have one translation. However, in the case of the Dutch and Italian wordnet there are many cases with multiple translations. For Subset1 we only considered the ILI-records occurring in the chains, for Subset2 we included all the ILI-records. This implies, that the increase in overlap is slightly less because the figures in D014D015 are too conservative (not taking all the equivalences into account).

Table 42 Coverage of ILI records (total)

	Total	(74995)	(42348)
	frequency	% $\cup$ (WN, IT, NL, ES)	% $\cup$ (IT, NL, ES)
ES	23201	30.94%	69.78%
IT	8670	11.56%	26.08%
NL	21465	28.62%	64.56%
$\cap$ (ES, IT)	6007	8.01%	18.07%
$\cap$ (ES, NL)	12561	16.75%	37.78%
$\cap$ (IT, NL)	5863	7.82%	17.63%
$\cap$ (ES, IT, NL)	4393	5.86%	13.21%

The next tables account for the coverage of complete chains (at node and edge level) for nouns and verbs, projected over WN1.5. Projections over the other wordnets are listed in Appendix II. As was pointed out before (see also D014D015) the figures presented in these tables are of rather limited use, since full coverage of the chains is rather difficult. The coverage in terms of complete chains is extremely low and the reason are: 1) the great differences in size between the different wordnets and 2) the fact that WN1.5 is an almost complete ontology with extremely long chains going from very general concepts to very specific ones (terminologic in some cases) while EWN subset2 is an ontology in construction, being far from complete and covering (according to our strategy) basically the most general concepts of the different languages. Consider, for instance, the overlapping between WN1.5 and the Spanish wordnet. The ratio, when comparing nodes is for nouns 28.20%. When comparing full chains this figure drops to 16.61%. This is not necessarily a bad result. A possible interpretation could be that most of the coverage is concentrated in the highest levels of the hierarchy. This is confirmed by other evidence. A more surprising result is the low (1.94%, only 1038 chains) coverage of complete WN1.5 chains by Dutch chains (dropping from 28.55 to 1.94). One reason is, of course, that Spanish WN has been built following as much as possible WN1.5 and Dutch WN in a more independent way, but another reason (that is related to the problem of long chains discussed above) is the fact that in some high levels of the hierarchy Dutch taxonomic relations are clearly different from English ones (which, in some cases, is a matter of choice).

Table 43 Coverage of complete noun chains projected over WN1.5 structure

	nodes	(53467)	edges	(53467)
	frequency	%	frequency	%
ES	8880	16.61	8880	16.61
IT	384	0.72	19	0.04
NL	1038	1.94	14	0.03
$\cap$ (ES, IT)	245	0.46	11	0.02
$\cap$ (ES, NL)	502	0.94	6	0.01
$\cap$ (IT, NL)	71	0.13	0	0.00
$\cap$ (ES, IT, NL)	52	0.10	0	0.00

Table 44 Coverage of complete verb chains projected over WN1.5 structure

	nodes	(8486)	edges	(8486)
	frequency	%	frequency	%
ES	1511	17.81	1511	17.81
IT	216	2.55	64	0.75
NL	1341	15.80	255	3.00
$\cap$ (ES, IT)	92	1.08	33	0.39
$\cap$ (ES, NL)	420	4.95	84	0.99
$\cap$ (IT, NL)	77	0.91	15	0.18
$\cap$ (ES, IT, NL)	38	0.45	7	0.08

We are mainly concerned with coverage of the more general subchains of WN1.5 rather than the complete chains. Obviously better, and more useful, results will be obtained when dealing with incomplete sequences (both subsequences and sequences containing gaps). These cases will be considered next. The following four tables account for the overlap of partial chains (node vs. edge, noun vs. verb) projected over WN1.5 structure, for different lengths of the chains.

Table 45 Coverage of partial noun chains of NODES projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	53467	53452	52575	52345	53452	51310	51270	53467
2	52928	35518	41122	39212	34357	25301	24614	53467
3	44677	23343	22597	21426	22082	13595	13568	53434
4	40559	17483	15557	14748	16894	8219	8203	52913
5	33260	11545	6815	5899	11286	2849	2826	50693
6	24421	6257	3018	2367	5960	1488	1476	45029
7	12842	1973	1289	889	1859	466	462	32299
8	5705	701	647	449	648	260	257	20558
9	2374	135	201	121	113	36	32	11821
10	961	21	30	12	16	3	2	5881
11	315	2	0	0	2	0	0	2576
12	130	0	0	0	0	0	0	1176
13	41	0	0	0	0	0	0	659
14	5	0	0	0	0	0	0	295
15	2	0	0	0	0	0	0	82

Table 46 Coverage of partial noun chains of EDGES projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	52928	25942	24676	23205	25073	14631	14631	53467
2	44677	7842	12937	12568	7656	7459	7459	53434
3	40559	1133	334	208	1122	51	51	52913
4	33260	42	48	11	28	0	0	50693
5	24421	0	0	0	0	0	0	45029
6	12842	0	0	0	0	0	0	32299
7	5705	0	0	0	0	0	0	20558
8	2374	0	0	0	0	0	0	11821
9	961	0	0	0	0	0	0	5881
10	315	0	0	0	0	0	0	2576
11	130	0	0	0	0	0	0	1176
12	41	0	0	0	0	0	0	659
13	5	0	0	0	0	0	0	295
14	2	0	0	0	0	0	0	82

Table 47 Coverage of partial VERB chains of NODES projected over WN1.5 structure

<i>LENGT H</i>	<i>ES</i>	<i>IT</i>	<i>NL</i>	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	<i>WN</i>
1	7906	7111	7986	7405	7018	6538	6456	8486
2	5644	3176	4932	3659	2925	1975	1817	8250
3	3302	1277	2148	1367	1161	477	420	6383
4	1561	608	722	371	560	66	44	3853
5	644	186	130	24	169	0	0	1894
6	198	25	15	0	10	0	0	865
7	54	0	5	0	0	0	0	403
8	13	0	0	0	0	0	0	153
9	1	0	0	0	0	0	0	44
10	1	0	0	0	0	0	0	12

Table 48 Coverage of partial VERB chains of EDGES projected over WN1.5 structure

<i>LENGTH</i>	<i>ES</i>	<i>IT</i>	<i>NL</i>	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	<i>WN</i>
1	5644	1590	2275	1870	1504	522	502	8250
2	3302	208	333	260	202	3	2	6383
3	1561	18	15	5	17	0	0	3853
4	644	0	2	0	0	0	0	1894
5	198	0	0	0	0	0	0	865
6	54	0	0	0	0	0	0	403
7	13	0	0	0	0	0	0	153
8	1	0	0	0	0	0	0	44
9	1	0	0	0	0	0	0	12

The corresponding tables for subset1 were used as an additional criterion for the building of subset2. If we consider the 8<sup>th</sup> column (intersection of Spanish, Italian and Dutch partial chains covering WN1.5 chains) the results are clearly better. The comparison of noun chains (NODES) can be seen in the following table. The figures show a clear improvement of the coverage.

Table 49 Comparison of partial coverage of WN1.5 chains by the intersection of WNs between subset1 and subset2

<i>Length</i>	<i>intersection subset1</i>	<i>intersection subset2</i>	<i>increment</i>	<i>% increment</i>
1	30909	51270	20361	66
2	16151	24614	8463	52
3	6756	13568	6812	100
4	2001	8203	6202	310
5	780	2826	2046	262
6	393	1476	1083	275
7	228	462	234	103
8	9	257	248	275
9	0	32		
10	0	2		
11	0	0		
12	0	0		
13	0	0		
14		0		
15		0		

For verbs the results for lengths 1, 2 and 3 increased from 5,382 to 6,456, from 765 to 1,817 and from 43 to 420. Additionally 44 chains of length 4 have been recorded.

The following tables account for the overlapping of partial chains containing one gap (node vs. edge, noun vs. verb) projected over WN1.5 for different lengths of the chain. Appendix III gives the projections over the Dutch, Italian and Spanish WN structure.

Table 50 Coverage of partial noun chains of NODES with 1 gap projected over WN1.5 structure

LENGT H	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
3	9607	5253	23133	21396	5707	9273	8837	53467
4	9472	5185	19252	17215	4921	6889	6444	53467
5	8088	3501	13551	12280	3064	5344	5095	53434
6	6686	2633	9244	8261	2152	3050	2883	52913
7	4916	1753	4325	3418	1398	711	622	50693
8	3326	843	1619	1013	664	282	238	45029
9	1380	328	448	268	277	68	65	32299
10	517	92	105	41	76	20	19	20558
11	147	11	11	9	10	3	3	11821
12	54	3	0	0	2	0	0	5881
13	16	0	0	0	0	0	0	2576
14	6	0	0	0	0	0	0	1176

Table 51 Coverage of partial NOUN chains of EDGES with 1 gap projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
3	0	3006	3092	3025	3005	766	765	53467
4	0	478	1233	1163	476	263	262	53434
5	0	73	223	126	70	50	50	52913
6	0	4	32	2	2	0	0	50693
7	0	0	5	0	0	0	0	45029

Table 52 Coverage of partial VERB chains of NODES with 1 gap projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
3	594	821	1361	992	702	635	591	8486
4	341	196	781	465	136	116	105	8250
5	206	72	317	153	55	5	5	6383
6	108	37	104	37	28	0	0	3853
7	60	3	29	5	2	0	0	1894
8	20	0	5	0	0	0	0	865
9	5	0	0	0	0	0	0	403

Table 53 Coverage of partial VERB chains of EDGES with 1 gap projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
3	0	148	42	20	147	0	0	8250
4	0	4	5	1	4	0	0	6383

Some interesting results can be pointed out. First of all, there is a clear improvement in the coverage of the intersection (column 8<sup>th</sup> of the tables). For nouns, for instance, and for NODE chains with one gap for lengths 3 to 11, there has been a growth in the size of intersection: from 5,672 to 8,837 (length 3), from 4,127 to 6,444 (length 4), from 2,901 to 5,095 (length 5), from 227 to 2,883 (length 6), from 11 to 622 (length 7), from 3 to 238 (length 8) and from 3 to 65 (length 9), with the additional recording of chains of length 10 and 11. Similar results have been obtained for verbs. In this case for chains of length 3 the increase is from 317 to 591 and for chains of length 4 from 2 to 105.

Another interesting result is that this increment has been obtained while in some cases the number of chains with gaps has decreased for individual WNs due to the process of filling gaps and so transforming partial or complete chains with gaps into partial or complete chains without gaps.

Consider, for instance the figures for Spanish (NODE chains of nouns) where one of the criteria for extension has been filling gaps in lengthy chains. The results for short chains are more or less the same than before (from 9,553 to 9,607 for length 3 chains and 9,293 to 9,472 for length 4 ones), but for longer chains the situation is the inverse (decreasing from 8,742 down to 8,088 for length 5 chains, from 7,721 down to 6,686 for length 6 ones, from 5,831 down to 4,916 for length 7, etc. For other languages this phenomenon does not appear so clearly because the filling of gaps appears jointly with a great increment in the number of chains but it exists in the same way.

#### 4.4. Comparison with composite ILIs

A great effort has been made by SHE to restructure the ILI to avoid mismatches due to extreme sense-differentiation in WN1.5. The most important restructuring consists of building composite ILI records (clustering) that group closely related senses. The composite ILIs are derived in a variety of ways: manual inspection, auto-hyponymy, comparison with WN1.6, analysis of sisters, diathesis alternations (Levin), etc. The details of this restructuring are reported in (Peters 98). From the total amount of composite ILIs we have selected sense-groupings of the type generalization (3853 clusters involving 8553 WN1.5 synsets). These group senses that are either difficult to keep apart (and thus may cause mismatches in ILI-chains) or can be seen as productive specializations of more general meanings. We replaced all occurrences of more specific ILI-records by composite ILIs that group these senses. After that we repeated the comparison to measure the overlap. The substitution of original ILIs by composite ones has been carried out in WN1.5 and in the ILI chains provided by the different sites.

We first present the results of individual chains (before comparison). Tables 54 and 55 are the composite counterpart of tables 34 and 35.

Table 54 *ILI chains for nouns (composite ILIs)*

	<i>ILI nodes</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>CHAINS</i>
<b>WN15</b>	59425	11	46270	13204	60151	52608
<b>ES</b>	19476	11	14854	4639	19887	17538
<b>NL</b>	19920	237	15980	4559	23385	37631
<b>IT</b>	4047	22	3820	245	6274	11299

Table 55 *ILI chains for verbs (composite ILIs)*

	<i>ILI nodes</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>CHAINS</i>
<b>WN15</b>	9716	547	7559	2197	9972	7891
<b>ES</b>	2615	321	1973	680	2873	2324
<b>NL</b>	5869	2	4727	1712	8558	9394
<b>IT</b>	1286	56	1110	159	2079	2285

There has been a small decrement in the number of ILI nodes (and in the number of chains before the process) as a logical result of the substitution of simple ILIs by composite ones. In the case of WN1.5 the decrement (from 60,557 to 59,425 for nouns) has been of 1.9%. For Spanish the decrement is similar (19,800 to 19,476, i.e. 1.6%) and also for Italian (4,111 to 4,047, i.e. 1.5%). In the case of Dutch (20,050 to 19,920, i.e. 0.6%) the decrement is smaller. This means that a relatively high amount of composite ILIs do not overlap (are not covered) by Dutch WNs.

The results for verbs are the following: In the case of WN1.5 the decrement (from 11,363 to 9716) has been of 15%. For Spanish the decrement is higher (3,507 to 2,615, i.e. 25%) and the effect of Italian (1,391 to 1,286, i.e. 8%) and Dutch (6,382 to 5,869, i.e. 8%) is smaller. These results could be explained by the different strategies followed for Spanish on the one hand (starting from

WordNet1.5 structure) and for Italian and Dutch on the other hand (starting with independent data). Furthermore, most clusters occur at the highest levels of the verb hierarchy, which is where most Spanish verbs are located. The fact that the decrement in size for WN1.5 is about 5 times greater in the case of verbs than in the case of nouns is due to the fact that the number of clustered synsets much larger: 2,646 (on a total of 60,557) for nouns and 5,904 (on a total of 11,363) for verbs.

Table 56 Frequencies and ratios of nominal chains / length /language (composite ILI)

	WN		ES		NL		IT	
	frequency	%	frequency	%	frequency	%	frequency	%
<b>1</b>					139	0.11	2	0.01
<b>2</b>	47	0.06	100	0.42	60	0.05	1365	10.00
<b>3</b>	681	0.93	824	3.45	1047	0.82	1831	13.41
<b>4</b>	2489	3.41	1940	8.11	5678	4.42	1289	9.44
<b>5</b>	7362	10.08	3573	14.94	14873	11.58	2628	19.25
<b>6</b>	15039	20.60	5340	22.33	17094	13.31	2428	17.78
<b>7</b>	14527	19.89	5284	22.10	20579	16.02	738	5.41
<b>8</b>	11614	15.91	3282	13.73	19536	15.21	1651	12.09
<b>9</b>	8412	11.52	1724	7.21	17907	13.94	1702	12.47
<b>10</b>	5538	7.58	773	3.23	13741	10.70	18	0.13
<b>11</b>	2920	4.00	352	1.47	9365	7.29		
<b>12</b>	1702	2.33	310	1.30	4764	3.71		
<b>13</b>	1263	1.73	245	1.02	1983	1.54		
<b>14</b>	732	1.00	107	0.45	961	0.75		
<b>15</b>	471	0.65	52	0.22	498	0.39		
<b>16</b>	210	0.29	2	0.01	166	0.13		
<b>17</b>	11	0.02	1	0.00	45	0.04		
<b>18</b>	2	0.00			5	0.00		
<b>19</b>					3	0.00		
<b>Total</b>	73020	100	23909	100	128444	100	13652	100
<b>Average</b>	7.60		6.69		7.88		5.49	

Table 57 Frequencies and ratios of verbal chains / length /language (composite ILI)

	WN		ES		NL		IT	
	frequency	%	frequency	%	frequency	%	frequency	%
<b>1</b>	207	1.46	107	3.22			1	0.03
<b>2</b>	1437	10.12	535	16.08			584	16.78
<b>3</b>	2498	17.60	777	23.35			869	24.97
<b>4</b>	2677	18.86	647	19.44			600	17.24
<b>5</b>	2355	16.59	513	15.41			654	18.79
<b>6</b>	1600	11.27	261	7.84			407	11.70
<b>7</b>	1210	8.53	193	5.80			244	7.01
<b>8</b>	798	5.62	132	3.97			99	2.84
<b>9</b>	514	3.62	103	3.09			22	0.63
<b>10</b>	479	3.37	44	1.32				
<b>11</b>	258	1.82	13	0.39				
<b>12</b>	99	0.70	3	0.09				
<b>13</b>	48	0.34						
<b>14</b>	11	0.08						
<b>15</b>	2	0.01						
<b>Total</b>	14193	100	3328	100			3480	100
<b>Average</b>	5.07		4.26				4.19	

Tables 56 and 57 (that have to be compared with Tables 38 and 39 for the non-composite case) present the distribution of length for the different chains corresponding to the different WNs. Some comments can be made. Consider the noun in WN1.5 in Table 55. If we compare this column with the corresponding one in Table 38 some surprising (at first glance) figures appear. In Table 38, the number of chains (total row) was 53,813. This figure is consistent with the corresponding Table 34. In the case of composite ILI, Table 56, the number of chains is 73,020. This figure is clearly higher than the figure in Table 54 (52,608). The average length of the chains in the composite case (7.60) is greater than the average length of the chains in the non-composite case (7.19). Moreover, the longest chain in the non-composite case was 16 nodes long while now there are 2 chains of length 18. Apparently, the substitution of simple synsets by composite ILIs creates additional overlapping between chains so that short chains concatenate into a longer one.

We will illustrate this with an example. In WN1.5, after substitution, the following 5 chains occur (clusters are identified by numbers beginning by 9):

- 90003222 90003059 90005400 90002166 90005945 00524290 90003305 00533007
- 90003218 90003305 90003213 90003086 90003087 90003309
- 90003218 90003305 90003309 90003304
- 90003321 90003304 90002093 90006266
- 90003134 90006266 00300916

As a result of this overlapping the following chains can be extracted:

- 90003222 90003059 90005400 90002166 90005945 00524290 90003305 00533007
- 90003218 90003305 00533007
- 90003222 90003059 90005400 90002166 90005945 00524290 90003305 90003213 90003086 90003087 90003309 90003304 90002093 90006266 00300916
- 90003218 90003305 90003213 90003086 90003087 90003309 90003304 90002093 90006266 00300916
- 90003222 90003059 90005400 90002166 90005945 00524290 90003305 90003309 90003304 90002093 90006266 00300916
- 90003218 90003305 90003309 90003304 90002093 90006266 00300916
- 90003321 90003304 90002093 90006266 00300916
- 90003134 90006266 00300916

From 5 chains with a maximum length of 8 we have obtained 8 chains with a maximum length of 15. Figure 2 shows the overlapping between these chains graphically (without the introduction of composite ILIs no overlap at all was produced). We have added a synset variant to the numbers to clarify what is going on. Composite ILIs (numbers preceded by 9) have multiple variants, representing the synsets which are collapsed. As we can see, the Composite hierarchy does not always make sense. Some groupings are acceptable, e.g. from 900033087 (register, file, register) to 90003213 (conserve, conserve) we get a single branch of 3 Composite ILIs that makes sense. Even the remainder of the chain, up to *communicate* and *act together* still makes some sense, although the relation between some of these nodes should not be hyponymy but some other type of relation, e.g. cause, role or subevent. A more doubtful case is represented by the Composite ILI 90003304 (effuse, breathe, keep, hold\_open). Below Figure 2, we listed the synsets from which this Composite ILI originates and the original WordNet1.5 hierarchy that goes with the synset members. The senses refer to rather different kinds of events and thus generate disjoint hyperonym chains as well. Apparently, some Composite ILIs at crucial positions in the hierarchy collapse incompatible senses (considering the further hyponymy relations). Strictly speaking, these Composite ILIs should not be of the polysemy type generalization but are metonymic groupings. A critical review of these complex Composite Chains may lead to a revision of the polysemy types which are now often extracted automatically.

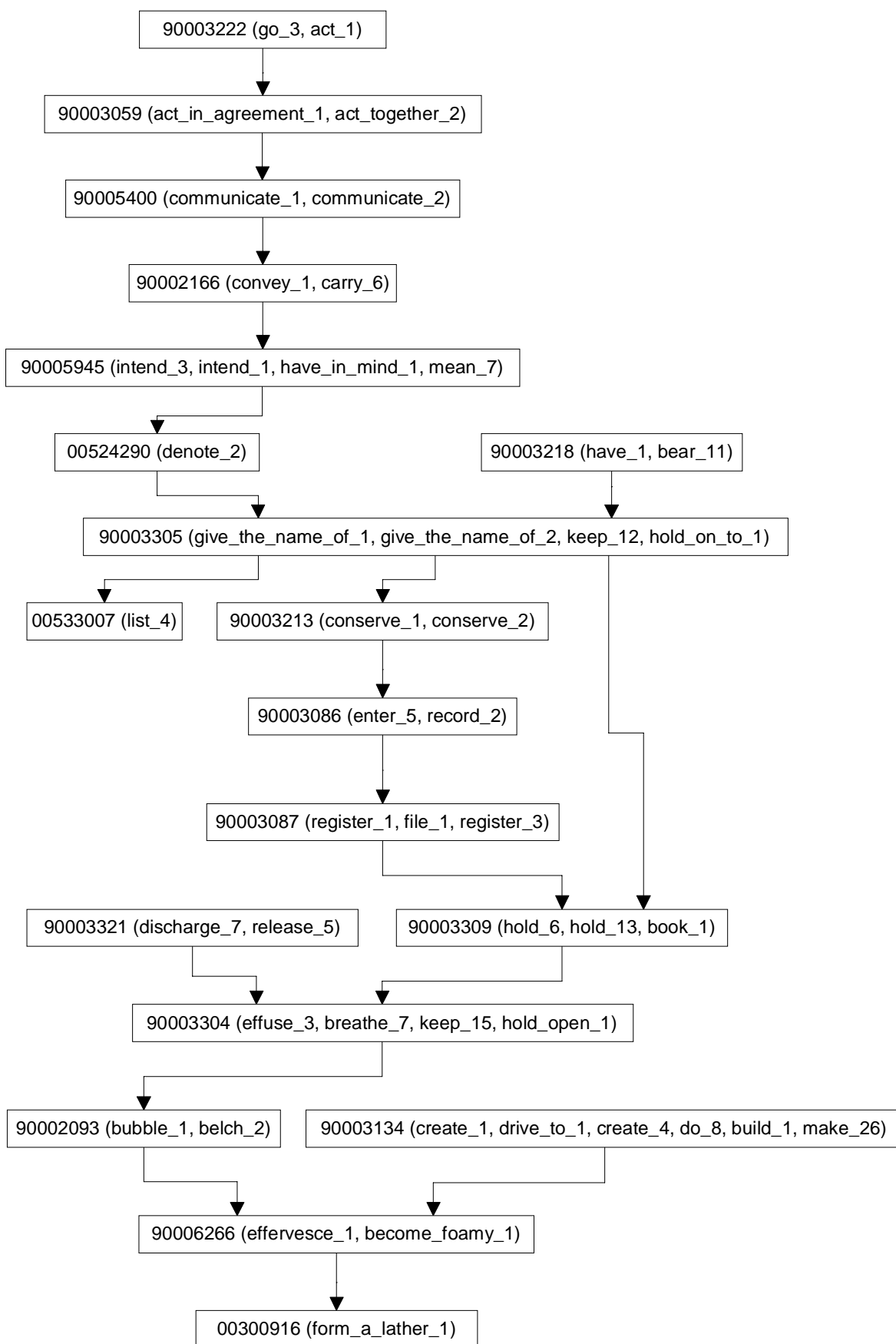


Figure 2. Overlapping between chains with composite ILLs

WordNet1.5 Synset: {effuse 1; give off#2}

WordNet1.5 File off-set: 175220

WordNet1.5 Gloss: "The room effuses happiness"

WordNet1.5 Hyperonyms: radiate#2

WordNet1.5 Synset: {give off 1; breathe#3; emit#1; pass off#1}

WordNet1.5 File off-set: 62533

WordNet1.5 Gloss: as of gases and odors

WordNet1.5 Hyperonyms: discharge#10

WordNet1.5 Synset: {keep 5; maintain#5}

WordNet1.5 File off-set: 604635

WordNet1.5 Gloss: maintain by writing regular records; "keep a diary"; "maintain a record"; "keep notes"

WordNet1.5 Hyperonyms: enter#1

WordNet1.5 Synset: {save 6; hold open#1; keep#9; keep open#1}

WordNet1.5 File off-set: 1299073

WordNet1.5 Gloss: as of a job or a seat

WordNet1.5 Hyperonyms: hold#21

discharge 10 breathe 3

discharge 10 breathe 3 radiate 2 give off 2

discharge 10 breathe 3 radiate 2 effuse 1

discharge 10 emit 1

discharge 10 give off 1

discharge 10 pass off 1

have 12 hold on to 2 hold 21 hold open 1

have 12 hold on to 2 conserve 2 enter 1 keep 5

have 12 hold on to 2 hold 21 keep 9

have 12 hold on to 2 hold 21 keep open 1

have 12 hold on to 2 conserve 2 enter 1 maintain 5

have 12 hold on to 2 hold 21 save 6

The results of coverage for complete chains for nouns are presented in Table 58. The results are (except for Italian node chains) a bit lower than the corresponding non-composite figures. Obviously the reason is the great increment of WN1.5 chains (from 52,762 to 90,404). These figures, however, as has been explained before have no special significance.

Table 58 Coverage of complete noun chains projected over WN1.5 structure (composite ILI)

	<i>nodes</i>		<i>edges</i>	
	<i>frequency</i>	<i>(73020)</i>	<i>frequency</i>	<i>(73020)</i>
		<i>%</i>		<i>%</i>
<b>ES</b>	11554	15.82	11346	15.54
<b>NL</b>	1356	1.86	12	0.02
<b>IT</b>	523	0.72	19	0.03
$\cap$ (ES, NL)	680	0.93	6	0.01
$\cap$ (ES, IT)	333	0.46	11	0.02
$\cap$ (IT, NL)	99	0.14	0	0.00
$\cap$ (ES, IT, NL)	77	0.11	0	0.00

Table 59 Coverage of complete verb chains projected over WN1.5 structure (composite ILI)

	<i>nodes</i>		<i>edges</i>	
	<i>frequency</i>	<i>(14193)</i> %	<i>frequency</i>	<i>(13986)</i> %
<b>ES</b>	1875	13.21	1579	11.29
<b>NL</b>	2147	15.13	211	1.51
<b>IT</b>	244	1.72	54	0.39
$\cap$ ( <b>ES, NL</b> )	585	4.12	66	0.47
$\cap$ ( <b>ES, IT</b> )	117	0.82	28	0.20
$\cap$ ( <b>IT, NL</b> )	110	0.78	14	0.10
$\cap$ ( <b>ES, IT, NL</b> )	62	0.44	7	0.05

The results for partial chains are presented in Tables 60 and 61 for nouns, and in Tables 62 and 63 for verbs. These tables have to be compared with Tables 45 and 46 for non-composite ILIs. If we restrict ourselves to node chains of nouns and compare the coverage of the intersection (IT+NL+SP), i.e. 8<sup>th</sup> column, there has been a clear improvement. For chains of length 1, the prior coverage (51,270 of 53,467, i.e. 96%) is more or less the same (70,855 of 73,020, i.e. 97%). For chains of length 2 the figures for non-composite ILIs are 24,614 of 53,467 (46%) and they are now 40,884 of 73,020 (i.e. 56%). The increment is in this case highly significant. For chains of length 3 the figures for non-composite ILIs are 13,568 of 53,434 (25%) and they are now 21,977 of 72,973 (30%). For chains of length 4 the figures for non-composite ILIs are 8,203 of 52,913 (15%) and for composite ILI they are now 11,626 of 72,292 (16%). The improvement for longer chains is thus less significant.

Table 60 Coverage of partial noun chains of NODES projected over WN1.5 structure (composite ILLs)

LENGTH	ES	NL	IT	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	73013	72104	72866	71883	72866	70905	70855	73020
2	72384	60170	55134	57203	53770	41579	40884	73020
3	63766	36836	38375	34554	36641	22010	21977	72973
4	57868	22250	29779	20450	28667	11646	11626	72292
5	47615	10117	20714	8494	18722	4187	4163	69803
6	34139	4811	12960	3900	11089	2373	2361	62441
7	20046	2138	3624	1600	2714	633	623	47402
8	10515	950	1404	660	928	271	268	32875
9	5542	281	446	167	278	70	65	21261
10	3410	45	200	22	64	4	3	12849
11	2048	1	50	1	2			7311
12	1388							4391
13	697							2689
14	413							1426
15	48							694
16	2							223
17	1							13

Table 61 Coverage of partial noun chains of EDGES projected over WN1.5 structure (composite ILLs)

LENGTH	ES	NL	IT	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	72382	36897	39702	34802	37467	21841	21838	73020
2	63744	16412	9877	15707	9624	9417	9417	72973
3	57795	500	1092	264	1080	6	6	72292
4	47409	73	38	11	26			69803
5	33826							62441
6	19643							47402
7	10270							32875
8	5398							21261
9	3324							12849
10	1997							7311
11	1376							4391
12	695							2689
13	413							1426
14	48							694
15	2							223
16	1							13

The next tables give the results for verbs. For chains of length 1, the figures for non-composite ILLs are 6,456 of 8,486 (76%) and for composite ILLs they are 12,285 of 14,193 (86%). For chains of length 2, the non-composite figures are 1,817 of 8,250 (22%) and the composite figures are 6,389 of 13,986 (i.e. 46%). For chains of length 3, the non-composite figures are 420 of 6,383 (7%) and the composite figures are 3,020 of 12,549 (24%). For chains of length 4, the non-composite figures are 44 of 3,853 (1%) and the composite figure 774 of 10,051 (7%). The increment in all cases is highly significant.

Table 62 Coverage of partial verb chains of NODES projected over WNI.5 structure (composite ILLs)

LENGTH	ES	NL	IT	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	13608	13838	12776	13327	12663	12368	12285	14193
2	11589	11403	8112	9956	7582	6754	6389	13986
3	8782	8020	4652	6368	4038	3357	3022	12549
4	5971	5187	2289	3493	1878	954	774	10051
5	3577	3109	953	1462	914	171	165	7374
6	1983	1729	398	569	389	38	38	5019
7	1080	933	143	278	137			3419
8	492	395	8	131	1			2209
9	197	147		62				1411
10	77	63		32				897
11	27	23		7				418
12	6	3						160

Table 63 Coverage of partial verb chains of EDGES projected over WNI.5 structure (composite ILLs)

LENGTH	ES	NL	IT	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	11354	5409	2957	4335	2830	972	946	13986
2	8302	592	354	418	340	2	2	12549
3	5306	24	17	9	17			10051
4	3200	2						7374
5	1668							5019
6	889							3419
7	411							2209
8	184							1411
9	65							897
10	17							418
11	6							160

The last results (Tables 64, 65, 66, and 67) represent the coverage of partial chains with one gap for nouns and verbs and for nodes and edges. The corresponding tables for non-composite ILI chains are 49, 50, 51 and 52. As is the case for partial chains without gaps the results here are very good and a clear improvement can be seen.

As before, we restrict ourselves to node chains of nouns and compare the coverage of the intersection (IT+NL+SP), i.e. 8<sup>th</sup> column. For chains of length 3, the coverage of the intersection for non-composite ILI was 8,837 of 53,467 (17%). For composite ILIs the coverage increased: 17,882 of 72,973 (25%). For length 4 chains, the non-composite coverage is 6,444 of 53,467 (12%) and the composite coverage is 13,417 of 72,292 (19%). For length 5 chains the coverage was 5,095 of 53,434 (10%) and now 10,237 of 69,803 (15%). For length 6 chains, the non-composite results are 2,883 of 52,913 (5%) and the composite results are 6,281 of 62,441 (11%). Even for chains of length 7 and 8 we see a positive result.

Table 64 Coverage of partial noun chains of NODES with 1 gap projected over WN1.5 structure (composite ILIs)

LENGTH	ES	NL	IT	$\cap$ (ES, NL)	$\cap$ (ES, IT)	$\cap$ (IT, NL)	$\cap$ (ES, IT, NL)	WN
3	14635	34079	13051	31768	14298	18454	17882	72973
4	14411	30006	12938	26780	13238	14017	13417	72292
5	12868	21960	9548	19845	9545	10582	10237	69803
6	11100	15770	7164	13408	7066	6512	6281	62441
7	8675	5981	5176	4672	4626	1077	964	47402
8	5609	2474	3631	1677	3275	516	471	32875
9	3006	658	542	396	332	55	50	21261
10	1441	137	150	56	78	17	16	12849
11	730	12	18	10	9	3	3	7311
12	455		3		2			4391
13	227							2689
14	135							1426
15	50							694
16	16							223
17	1							13

Table 65 Coverage of partial noun chains of EDGES with 1 gap projected over WN1.5 structure (composite ILIs)

LENGTH	ES	NL	IT	$\cap$ (ES, NL)	$\cap$ (ES, IT)	$\cap$ (IT, NL)	$\cap$ (ES, IT, NL)	WN
3	408	3469	6141	3355	6140	805	804	72292
4	406	1196	357	1120	353	182	181	69803
5	402	176	28	80	25	5	5	62441
6	319	31		2				47402
7	214	5						32875
8	135							21261
9	85							12849
10	42							7311
11	12							4391
12								2689
13								1426
14								694
15								223
16								13

Similar results are obtained for verbs. For length 3 chains, the coverage of the intersection for non-composite ILI is 591 of 8,486 (7%). For composite ILIs the coverage increased: 2,362 of 12,549 (19%). For length 4 chains, the non-composite coverage is 105 of 8,250 (1%) and the composite coverage is 1,389 of 10,051 (14%). For length 5 chains, the non-composite coverage is 5 of 6,383 (0%) and the composite coverage is 841 of 7,374 (11%). Additionally, 285 chains of length 6 and 45 chains of length 7 have been recorded.

Table 66 Coverage of partial verb chains of NODES with 1 gap projected over WN1.5 structure (composite ILIs)

LENGTH	ES	NL	IT	$\cap$ (ES, NL)	$\cap$ (ES, IT)	$\cap$ (IT, NL)	$\cap$ (ES, IT, NL)	WN
3	2193	2762	2639	3377	2664	2266	2362	12549
4	1908	2172	1694	2713	1596	1373	1389	10051
5	1575	1570	1253	2038	1059	903	841	7374
6	1189	1106	722	1296	512	415	285	5019
7	836	692	465	626	298	60	45	3419
8	624	400	309	235	177	2		2209
9	409	227	145	70	132			1411
10	212	111	12	29	8			897
11	76	55		10				418
12	19	22						160
13		4						61
14								13
15								2

Table 67 Coverage of partial verb chains of EDGES with 1 gap projected over WN1.5 structure (composite ILIs)

LENGTH	ES	NL	IT	$\cap$ (ES, NL)	$\cap$ (ES, IT)	$\cap$ (IT, NL)	$\cap$ (ES, IT, NL)	WN
3	95	237	596	138	593	12	10	10051
4	71	23	4	6	3			7374
5	45	2						5019
6	24							3419
7	8							2209

## 5. Conclusions

In general, we can conclude that the wordnets have progressed considerably compared to Subet1 and are reaching their final state. Quantitatively, we made the following progress:

- all wordnets have reached their final quantitative state: 23-25K synsets and 40-55K word meanings, which is about 1/3 of the size of WordNet1.5.
- all 3 wordnets include the most frequent entries from the corresponding Parole lexicons in the languages.
- the average number of relations for Dutch, Italian and Spanish in Subset2 (2.0 to 2.5) is more in balance as compared to Subset1 (1.5 to 2.8), and is slightly higher than WordNet1.5 (1.4).
- clustering across the top-ontology is balanced across WordNet1.5, Spanish and Dutch, except for Animal, Plant and Group.

Qualitative improvements with respect to Subet1 are:

- the number of synonyms or synset variants per synset has become more balanced across the wordnets: between 1.09 and 2.27 for Subet1 and between 1.21 and 1.65 in Subset2.
- the ILI-intersection for 3 languages has increased 7 times, and the intersection of ILIs for 2 languages represents up to 20 to 40% of all covered ILI-records.
- the length of the ILI chains got more balanced.
- the overlap of partial chains increased considerably: from 1.8 to 4 times for nouns (including chains of length 8!) and from 1.2 up to 10 times for verbs
- the overlap of partial chains with one gap increased even more considerably: up to 4 times for nouns and up to 60 times for verbs

We also investigated the effect of the Composite ILIs on the compatibility of the ILI-chains. We concluded that the Composite ILIs:

- raised the complexity of the chains
- increased the overlap of partial chains: up to 10% for nouns and up to 24% for verbs
- increased the overlap of partial chains with one gap: up to 7% for nouns and up to 13% for verbs

We also concluded that more work needs to be done with respect to:

- the quality of equivalence relations (this would e.g. eliminate pathologically long ILI-chains in Dutch).
- the number of equivalence relations in the Italian wordnet.

Finally, we stated that complete overlap is both impossible and not desirable. The size of the wordnets in EuroWordNet is 1/3 of the size of WordNet1.5, which simply makes it impossible to have full overlap of hierarchies. However, in many cases there are structural differences between the wordnets at high levels in the hierarchy. In some cases these differences are chosen on principled grounds (disagreeing with the WordNet1.5 classification or not applying to lexicalizations in other languages). A difference at a high level immediately implies that all chains below this node cannot be the same across the wordnets.

## References

- Atserias J., S. Climent, J. Farreres, G. Rigau, H. Rodriguez,  
 1997 *Combining Multiple Methods for the Automatic Construction of Multilingual WordNets*, Proceedings of Conference on Recent Advances on NLP. RANLP 97. Tzigov Chark, Bulgaria, 1997.
- Kruyt, T.  
 1998 "Electronische woordenboeken en tekstcorpora voor Europese taaltechnologie", Trefwoord, 12, 1997-1998, Sdu Uitgevers, Den Haag/ Antwerpen.
- Miller G.A, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller,  
 1990 *Introduction to WordNet: An On-line Lexical Database*, In: International Journal of Lexicography, Vol. 3, No.4, 235-244.
- Peters W.  
 1998 "Restructured ILI" EuroWordNet (LE 4003), Deliverable 2D004, University of Sheffield.
- Rodriguez, 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, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, W. Peters.  
 1997 The EuroWordNet Base Concepts and Top Ontology. EuroWordNet (LE 4003) Deliverable D017, D034, D036. University of Amsterdam
- Vossen, P., L. Bloksma, S. Climent, M.A. Marti, G. Oreggioni, G. Escudero, G. Rigau, H. Rodriguez, C. Peters, A. Roventini, F. Bertagna, A. Alonge, W. Peters.  
 1998 The Restructured Core wordnets in EuroWordNet: Subset1. EuroWordNet (LE 4003), Deliverable D014D015, University of Amsterdam.
- Vossen, P.  
 1998 Vossen, P. Introduction to EuroWordNet. In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds), Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3 1998. 73-89.

## Appendix I: Explanations of Gaps in the Dutch wordnet compared to Spanish, Italian and English.

The next examples illustrate the differences between the Dutch hierarchy and WordNet1.5 that have been found when inspecting the most important GAPS in the Dutch hierarchy, as compared to the other wordnets:

```
--- # 'relation' 00017862 EQ_S met 'relatie_1/betrekking_2'
```

Difference in classification:

Hyperonym in WN1.5 is 'abstraction'

Hyperonym in DWN is 'toestand' (state\_1)

We do not have the equivalent for 'abstraction' in our database (which should be there), however we do not agree with the classification either.

```
--- # 'social relation' 00018392 Eq_has_hyperonym = 'relatie_1/betrekking_2' (relation)
```

No equivalent in Dutch for 'social relation' (would be a multi word). Again difference in hierarchy: in WN1.5 via 'relation' up to 'abstraction'. In Dutch to 'state' and not to 'abstraction'.

```
--- # 'communication' 00018599 Eq_has_hyponym 'betekenis_1' (meaning)
```

Difference in classification:

'communication' goes to 'social relation' -> 'relation' -> 'abstraction'

betekenis\_1 (meaning) goes to eigenschap (attribute).

An equivalent for 'communication' in Dutch in this sense does exist, but is not in our database. Therefore it is linked by a hyponym-equivalent link.

```
--- # 'food/nutrient' 00011263 EQ_N_S 'voedsel'
```

Difference in hierarchy:

Hyper of 'food' in WN is 'substance, matter'

The Hyperonym of the Dutch equivalent 'voedsel' was the Top node 'iets' (anything). We agree that it should be assigned to 'substantie' (substance), therefore we changed it.

```
--- # administrative district \# 1 Eq_has_hyperonym district_1
```

GAP: would be multi word in Dutch

```
--- # content/ message/ subject matter/ substance 04313427-n
```

Had no equivalent yet in Dutch. We assigned inhoud\_3 (content) as the closest Dutch synset to this ILI, but the hyperonym in WN is still 'communication', whereas in Dutch we still prefer betekenis (meaning) as HYPER.

```
--- # psychological feature 00012517-n Eq_has_hyperonym eigenschap (attribute)
```

GAP: would be multiword in Dutch; hierarchy difference: in WN it is a top.

```
--- # definite quantity \# 1 Eq_has_hyperonym hoeveelheid (quantity)
```

GAP: would be multiword in Dutch, also difference in hierarchy.

in WN -> quantity -> abstraction

in DWN -> grootheid (quantity) -> maat (quantity/amount) -> eigenschap (attribute) -> iets

Hierarchy around this concept not fully worked out yet.

```
--- # mental object \# 1 Eq_has_hyponym gedachte (thought/idea)
```

GAP: would be multi word (artificial). Hierarchy differences:

WN -> cognition -> psychological feature

DWN -> voorstelling (opinion perspective) -> geestgesteldheid (cognitive state of mind)

-> gesteldheid (condition) -> toestand (state)

--- # geographic area \# 1 EQ\_N\_S gebied \# 2 (also area/ region)  
hierarchy difference:  
WN -> area -> region -> location -> space -> abstraction  
DWN -> plaats (location/ place) -> iets

## Appendix II Projection of complete chains on the Dutch, Italian and Spanish wordnets

Table 68 Coverage of complete noun chains projected over Dutch wordnet structure

	<i>nodes</i>		<i>edges</i>	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
<b>ES</b>	17747	19.32	58	0.06
<b>IT</b>	3	0.00	3	0.00
<b>∩ (ES, IT)</b>	2	0.00	2	0.00

Table 69 Coverage of complete verb chains projected over Dutch wordnet structure

	<i>nodes</i>		<i>edges</i>	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
<b>ES</b>	2565	4.43	2	0.00
<b>IT</b>	58	0.10	0	0.00
<b>∩ (ES, IT)</b>	36	0.06	0	0.00

Table 70 Coverage of complete noun chains projected over Italian wordnet structure

	<i>nodes</i>		<i>edges</i>	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
<b>ES</b>	5397	48.46	50	0.45
<b>NL</b>	1293	11.61	80	0.72
<b>∩ (ES, NL)</b>	1053	9.45	12	0.11

Table 71 Coverage of complete verb chains projected over Italian wordnet structure

	<i>nodes</i>		<i>edges</i>	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
<b>ES</b>	915	28.45	35	1.09
<b>NL</b>	296	9.20	33	1.03
<b>∩ (ES, NL)</b>	187	5.81	14	0.44

Table 72 Coverage of complete noun chains projected over Spanish wordnet structure

	<i>nodes</i>		<i>edges</i>	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
<b>NL</b>	1077	6.07	12	0.08
<b>IT</b>	548	3.09	18	0.10
<b>∩ (NL, IT)</b>	120	0.68	0	0.00

Table 73 Coverage of complete verb chains projected over Spanish wordnet structure

	<i>nodes</i>		<i>edges</i>	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
<b>NL</b>	786	30.64	198	7.72
<b>IT</b>	213	8.30	76	2.96
<b>∩ (NL, IT)</b>	93	3.63	29	1.13

### Appendix III Projection of partial chains on the Dutch, Italian and Spanish wordnets

Table 74 Coverage of partial noun chains of NODES with 1 gap projected over Dutch wordnet structure

LENGTH	ES	IT	$\cap$ (ES, IT)	NL
3	41194	19704	19562	91855
4	38609	17989	17361	91716
5	33305	14416	14025	91664
6	26319	10294	10144	90711
7	17806	5693	5598	85304
8	9169	2081	2034	71948
9	3482	390	374	57371
10	899	39	35	42282
11	140	0	0	29238
12	15	0	0	18099
13	4	0	0	9630
14	1	0	0	3872

Table 75 Coverage of partial noun chains of EDGES with 1 gap projected over Dutch wordnet structure

LENGTH	ES	IT	$\cap$ (ES, IT)	NL
3	1096	441	97	91716
4	251	135	7	91664
5	82	14	0	90711
6	11	0	0	85304

Table 76 Coverage of partial verb chains of NODES with 1 gap projected over Dutch wordnet structure

LENGTH	ES	IT	$\cap$ (ES, IT)	NL
3	27508	30953	30432	57836
4	26277	26239	25215	57836
5	23549	21416	20409	57796
6	19842	12543	11866	57414
7	16664	8173	7825	56325
8	13389	3403	3152	53164
9	10139	1137	950	47593
10	5233	302	226	39519
11	2684	8	2	31185
12	1444	0	0	23627
13	760	0	0	17319
14	393	0	0	11723
15	27	0	0	7345
16	3	0	0	4344

Table 77 Coverage of partial verb chains of EDGES with 1 gap projected over Dutch wordnet structure

LENGTH	ES	IT	$\cap$ (ES, IT)	NL
3	671	20	0	57836
4	61	0	0	57796
5	1	0	0	57414

Table 78 Coverage of partial noun chains of NODES with 1 gap projected over Italian wordnet structure

LENGTH	ES	NL	$\cap$ (ES, NL)	IT
3	36	2389	2209	11138
4	30	2122	1960	11136
5	19	1506	1363	9697
6	10	647	581	7821
7	9	382	370	6537
8	6	236	220	3794
9	0	48	40	1731

Table 79 Coverage of partial noun chains of EDGES with 1 gap projected over Italian wordnet structure

LENGTH	ES	NL	$\cap$ (ES, NL)	IT
3	1640	285	250	11136
4	525	150	18	9697
5	76	10	0	7821
6	10	0	0	6537

Table 80 Coverage of partial VERB chains of NODES with 1 gap projected over Italian wordnet structure

LENGTH	ES	NL	$\cap$ (ES, NL)	IT
3	855	1039	736	3216
4	725	859	553	3215
5	562	527	291	2374
6	328	210	92	1664
7	126	46	12	1177
8	23	4	0	759
9	1	0	0	387

Table 81 Coverage of partial VERB chains of EDGES with 1 gap projected over Italian wordnet structure

LENGTH	ES	NL	$\cap$ (ES, NL)	IT
3	13	2	0	3215
4	8	0	0	2374

Table 82 Coverage of partial noun chains of NODES with 1 gap projected over Spanish wordnet structure

LENGTH	NL	IT	$\cap$ (NL, IT)	ES
3	8104	1642	3348	17747
4	7011	1567	2573	17747
5	4938	1071	2056	17646
6	3190	728	1266	16885
7	1514	425	307	15094
8	548	258	130	12165
9	161	118	29	8108
10	25	38	10	4101
11	4	6	2	1726
12	0	1	0	662

*Table 83 Coverage of partial noun chains of EDGES with 1 gap projected over Spanish wordnet structure*

<i>LENGTH</i>	<i>NL</i>	<i>IT</i>	$\cap$ ( <i>NL, IT</i> )	<i>ES</i>
<b>3</b>	1108	1228	352	17747
<b>4</b>	481	187	112	17646
<b>5</b>	69	36	15	16885
<b>6</b>	5	5	3	15094

*Table 84 Coverage of partial VERB chains of NODES with 1 gap projected over Spanish wordnet structure*

<i>LENGTH</i>	<i>NL</i>	<i>IT</i>	$\cap$ ( <i>NL, IT</i> )	<i>ES</i>
<b>3</b>	378	251	226	2565
<b>4</b>	226	64	49	2385
<b>5</b>	89	18	3	1645
<b>6</b>	25	9	0	890
<b>7</b>	6	2	0	389

*Table 85 Coverage of partial VERB chains of EDGES with 1 gap projected over Spanish wordnet structure*

<i>LENGTH</i>	<i>NL</i>	<i>IT</i>	$\cap$ ( <i>NL, IT</i> )	<i>ES</i>
<b>3</b>	15	58	0	2385
<b>4</b>	1	4	0	1645