Validation of Meaning

WP8.1

Deliverable 8.1

http://www.lsi.upc.es/~nlp/meaning/meaning.html

IST-2001-34460

Piek Vossen

Elton Glaser

Hetty van Zutphen

Rachel Steenwijk

Irion Technologies BV

info@irion.nl

http://www.irion.nl

Table of Contents

1 Introduction

The goal of MEANING is to acquire large-scale lexical knowledge from corpora and the web that can be used to improve Word-Sense-Disambiguation. The knowledge is stored in a central multilingual repository (MCR), structured as the EuroWordNet database. The MCR uses an Inter-Lingual-Index (ILI) and holds wordnets for English, Italian, Spanish, Catalan and Basque. The type of knowledge acquired in MEANING is mostly conceptual and stored as relations and features for concepts. Likewise, it can be acquired in one language and ported to another language via the ILI.

Currently, MEANING is acquiring:

- Domain classification of word meanings, including named-entities
- Topic signatures
- Selectional restrictions
- Lexical relations from morphological structure of words
- Sense tagged examples for the involved languages

In addition to the acquisition of knowledge, MEANING also develops Word-Sense-Disambigation (WSD) technologies that can exploit this knowledge. WSD can help the acquisition of knowledge and the acquired knowledge will help WSD.

MEANING follows 3 major cycles. In the first phase of the project, the MCR is built and existing resources and data are uploaded into the MCR. In the meanwhile, WSD is used to sense-tag corpora and the acquisition of knowledge from corpora and the WWW starts. There is no use of MEANING results yet, hence we refer to MCR0, WSD0 and acquisition (ACQ0) in this cycle.

In the second cycle the acquired knowledge in the MCR is used to acquire more knowledge from corpora and the WWW and to apply further sense-tagging. This results in MCR1, WSD1 and ACQ1 at the end of the second cycle. The third cycle finally results in ACQ2 and WSD2 results that are stored in MCR2:

Cycle 1, month 1 - 12:

- MCR0: filled with existing knowledge and wordnets, that can be ported to other languages and systems.
- WSD0: disambiguation using existing knowledge resulting in sense-tagged corpora.
- ACQ0: acquisition of knowledge with current state-of-the-art technology.

Cycle 2, month 12 - 21:

- WSD1: improved disambiguation with new knowledge from ACQ0 and MCR0
- ACQ1: improved acquisition with WSD0 and MCR0
- MCR1: filled with the result of ACQ1 and WSD1

Cycle 3, month 21 - 30:

- WSD2: improved disambiguation with new knowledge from ACQ1 and MCR1
- ACQ2: improved acquisition with WSD1 and MCR1
- MCR2: filled with the result of ACQ2 and WSD2

Evaluation of the results is carried out at two different levels at the end of each cycle. The consortium members will carry out a separate benchmarking of the techniques using standard evaluation measures. For example, WSD is tested on Senseval data and with Senseval evaluation metrics. This work will be carried out in Work Package 7.

In addition to the benchmarking, Irion Technologies will carry out the validation of the results for integration and exploitation in commercial language-technology products (Work Package 8). This validation does not directly measure the correctness of the individual techniques but moreover measures the exploitation for end-user applications. The end-user applications should significantly benefit from the enabling technologies. This both implies that the performance of the end-user application is improved and that the effort of developing and integrating the enhancement is justifiable to customers.

The current integration of MEANING results relates to the first cycle, using MCR0 and WSD0. At the end of cycle 2 and cycle 3, there will be another integration and validation using MCR1 and MCR2 respectively. Furthermore, Irion has distributed the architecture for building indexes to the consortium so that they can apply WSD directly to the data that is used for testing. Data disambiguated by the MEANING partners can then be indexed and exploited by the Irion applications with the same test set. This means that in addition to the MCR itself also the derived WSD techniques are investigated.

The current integration and validation has been carried out using the Reuters test collection. In the future, we will also apply the same process to the EFE database. EFE has recently joined the project as a subcontractor and they have provided sample data for evaluation. These data have not yet been processed and the results will thus be reported in the next deliverable 8.2.

This deliverable (8.1) then describes the integration and a first evaluation of the MEANING results in two applications at Irion Technologies:

- Cross-lingual information retrieval
- Mono-lingual document classification

Word Sense Disambiguation (WSD) may be useful for these applications for three major reasons:

1. Increase precision
2. Reduce indexing and search time and effort
3. Provide cross-lingual functionality

To test the usefulness of WSD, we built indexes on the REUTERS news collection. Different indexes have been built with and without word-sense-disambiguation. We compared the results for a set of test queries and test documents across each index. The test queries for retrieval have been created manually from an automatically derived set. The test documents for classification are automatically extracted.

Before we can integrate the results of MEANING, we needed to make a mapping between the MEANING database and our own corporate database. The mapping is also described in this deliverable.

In the next section, we will first describe the applications. In section 3, we will describe the resources that are used. Section 4 explains our approach to WSD and how it is integrated in the applications. Section 5 describes the test corpus and section 6 and 7 describe the results for cross-lingual retrieval and mono-lingual document classification respectively.

2 Overall architecture of Irion applications

2.1 TwentyOne Search

Irion Technologies applies the following cascaded approach:

1. Document collection
2. Document conversion to XML
3. Concept extraction
4. Indexing

For the concept extraction, a series of NLP processes are applied:

1. Tokenization: detect textual and sentence boundaries;
2. Tagging: assign initial POS tags to words;
3. Shallow parsing: assign noun-phrase boundaries to phrases;
4. Named Entity Recognition: detect names and entity references within noun phrases;
5. Concept recognition: detect relevant meanings and expand to relevant synonyms and or translations;
6. Normalization: reduce the expansions to a root form;

Each of these processes generates output files in XML format that are used for the next process.

In the first phase, a division is made between the textual content and the meta information. The textual content is stored in a HTML file and the meta data on the document is represented in a so-called MET file. The MET file contains administrative information and meta information that can be used in the retrieval, e.g. titles, authors, source, type of document, date, topic classification, region, etc.. Users are free to specify the meta information.

Next, the first phase of morpho-syntactic analysis generates a so-called SYN file from the HTML file. The SYN file contains all the detected phrases in the HTML document (step 1, 2, 3). The SYN file is the input for a named entity recognition program. This program uses a semantic network of names and some heuristics to detect name phrases within the NPs of the SYN file. The NPs and Names are stored in the NAM file. The NAM file is then fed to the Concept Recognition program. This program uses a multilingual semantic network, called

SemNet, and it expands the content words in the NPs and the Names to all synonyms. It also generates translations for each NP and each Name in any of the specified target languages. The output of the Concept Recognition program is an XXX file. The program generates separate XXX files for each language. So an original HTML document that is taken as input leads to the generation of:

- MET file with meta information
- SYN file with NPs
- NAM file with NPs and Names
- XXX file with concepts in each language

From the XXX files, an index is built for each language. The retrieval then starts by identifying the language of the query. Knowing the query language, it calls the proper NLP library to process it and matches the query with the corresponding index. The index yields a pointer to a XXX document and NPs or Names within that document. These XXX documents correspond to the original HTML file that can be listed or shown, where the NPs are used for highlighting. The next figure gives an overview of this process for the cross-lingual retrieval application.

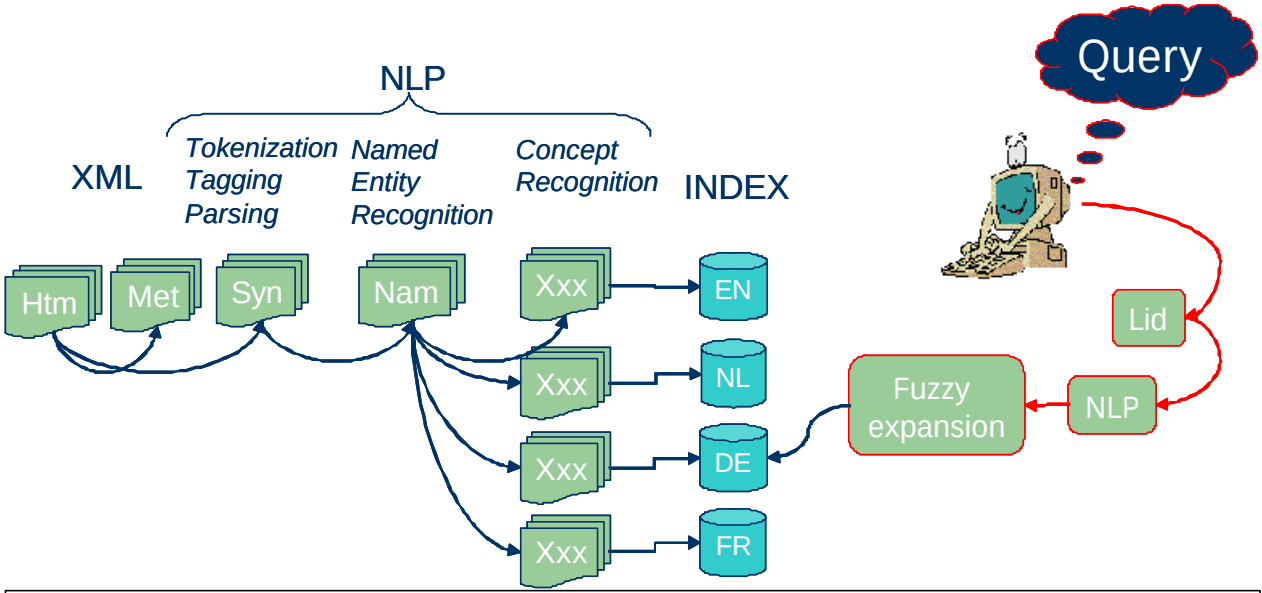

Figure 1: Overview of Concept extraction and Indexing

Concept recognition (step 5) is the most crucial process where the MEANING results can be useful. Currently, all content words in NPs that are not recognized as Named Entities are fully expanded to all its meanings. Obviously, the expansion leads to an enormous improvement of the recall, but it also results in very large indexes and introduces many false hits.

An XXX file consists of a list of NPs with the original phrase and the content words that have been extracted and matched with the semantic network. The NPs have identifiers that correspond to identifier tags that are inserted in the original HTML file from which they are extracted. In the XML encoding, a distinction is made between:

- <SS> = source strings
- <WRD> = lemmas in the semantic network
- <NWR> = normalized forms
- <CE> = compound element

Furthermore, multiword compounds (e.g. “police cell”) are retrieved in the semantic network as a single unit and singleword compounds that cannot be found are split into matched elements. Only the strings in the NWR tags are added to the index.

Below, you see a snapshot of such a file from the Reuters collection that has a full expansion to all associated concepts and all associated synonyms.

[Example of Full expansion for English, Spanish and Italian]

English (original)

```
<NP ID="13">
  <ORIG_PHRASE>futures contracts</ORIG_PHRASE>
  <SW><SS>futures</SS><NWR>futur</NWR>
    <WR>futures</WR><NWR>futur</NWR><WR>futures dealings</WR><NWR>futur
deal</NWR><WR>future</WR><NWR>futur</NWR><WR>forward transactions</WR><NWR>forward
transact</NWR><WR>future</WR><NWR>futur</NWR><WR>future tense</WR><NWR>futur tens</NWR></SW>
  <SW><SS>contracts</SS><NWR>contract</NWR><MW>economy; </MW>

<WR>treaty</WR><NWR>treati</NWR><WR>union</WR><NWR>union</NWR><WR>contract</WR><NWR>contract</
NWR><WR>bond</WR><NWR>bond</NWR><WR>condition</WR><NWR>condition</NWR><WR>pact</WR><NWR>pact</
NWR><WR>bargain</WR><NWR>bargain</NWR><WR>relationship</WR><NWR>relationship</NWR><WR>engageme
nt</WR><NWR>engag</NWR><WR>alliance</WR><NWR>allianc</NWR><WR>undertaking</WR><NWR>undertak</N
```

```
WR><WR>association</WR><NWR>associ</NWR><WR>contracting-out</WR><NWR>contracting-
out</NWR><WR>arrangement</WR><NWR>arrang</NWR><WR>settlement</WR><NWR>settlem</NWR><WR>composi
tion</WR><NWR>composit</NWR><WR>obligation</WR><NWR>oblig</NWR><WR>tender</WR><NWR>tender</NWR
><WR>commitment</WR><NWR>commit</NWR><WR>covenant</WR><NWR>coven</NWR><WR>deal</WR><NWR>deal</
NWR><WR>agreement</WR><NWR>agreem</NWR></SW>
</NP>
```

Spanish (translation)

```
<NP ID="13">
  <SW><SS>futures</SS>

<WR>porvenir</WR><NWR>porvenir</NWR><WR>mañana</WR><NWR>manana</NWR><WR>futuro</WR><NWR>futuro
</NWR><WR>operación</WR><NWR>operacion</NWR><WR>destino</WR><NWR>destino</NWR></SW>
  <SW><SS>contracts</SS><MW>economy; </MW>

<WR>tratado</WR><NWR>tratado</NWR><WR>convenio</WR><NWR>convenio</NWR><WR>concertación</WR><NW
R>concertacion</NWR><WR>convenio regulador</WR><NWR>convenio
regulador</NWR><WR>acuerdo</WR><NWR>acuerdo</NWR><WR>negocio</WR><NWR>negocio</NWR><WR>obligac
ión</WR><NWR>obligacion</NWR><WR>subasta</WR><NWR>subasta</NWR><WR>subasta
pública</WR><NWR>subasta
publica</NWR><WR>adjudicación</WR><NWR>adjudicacion</NWR><WR>contrata</WR><NWR>contrata</NWR><
WR>contrato</WR><NWR>contrato</NWR><WR>trato</WR><NWR>trato</NWR><WR>concurso</WR><NWR>concurs
o</NWR><WR>compromiso</WR><NWR>compromiso</NWR><WR>pacto</WR><NWR>pacto</NWR></SW>
</NP>
```

Italian (translation)

```
<NP ID="13">
  <SW><SS>futures</SS>

<WR>avvenire</WR><NWR>avvenire</NWR><WR>domani</WR><NWR>domani</NWR><WR>futuro</WR><NWR>futuro
</NWR><WR>destino</WR><NWR>destino</NWR></SW>
  <SW><SS>contracts</SS><MW>economy; </MW>

<WR>obbligazione</WR><NWR>obbligazione</NWR><WR>patto</WR><NWR>patto</NWR><WR>accordo</WR><NWR
>accordo</NWR><WR>legame</WR><NWR>legame</NWR><WR>unione</WR><NWR>unione</NWR><WR>appalto</WR>
<NWR>appalto</NWR><WR>contratto</WR><NWR>contratto</NWR><WR>contratto
d'appalto</WR><NWR>contratto d'appalto</NWR><WR>connubio</WR><NWR>connubio</NWR></SW>
</NP>
```

2.2 TwentyOne Classify

The second important product of Irion is the Classification engine. The classification engine allows you to train a classifier by giving it a set of documents with classes. The classifier, which is called a Monk, can then assign these classes to unseen documents. The next figure, gives an overview of this process.

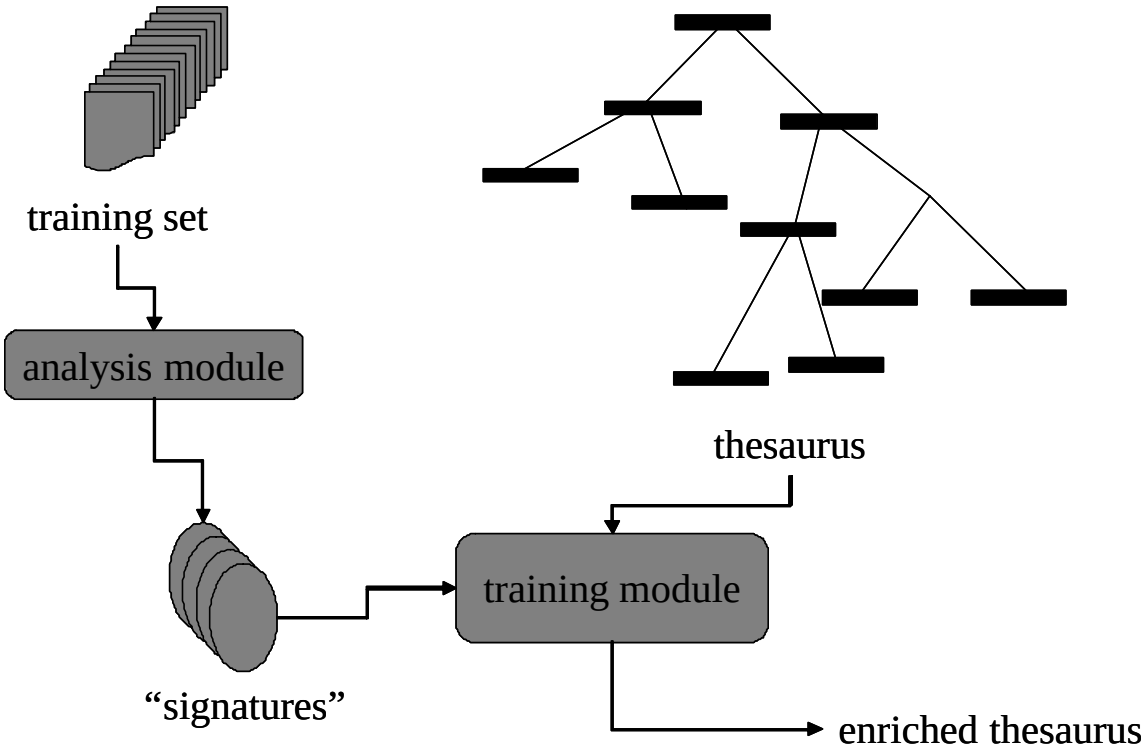

Figure 2: Overview of the training process for TwentyOne Classifcation

The training module extracts a so-called signature from the textual structure of the training document. This signature is associated with a thesaurus label. When classifying a document, it will compare the signature of the incoming document with the documents in the training set an extract a score for the categories of the most similar document. The thesaurus labels can be organized hierarchically and the system can assign more than one thesaurus label to a document. The system provides many options and tools to evaluate the quality of the classifier and to give feedback and suggestions to improve it.

A trained classification system is called a Monk and can be kept in a so-called Monastery. You can make as many monks as you like each with different classification behaviours. You

are free to deploy these monks in any classification environment. There is a general API layer that can be called by any program to classify any documents.

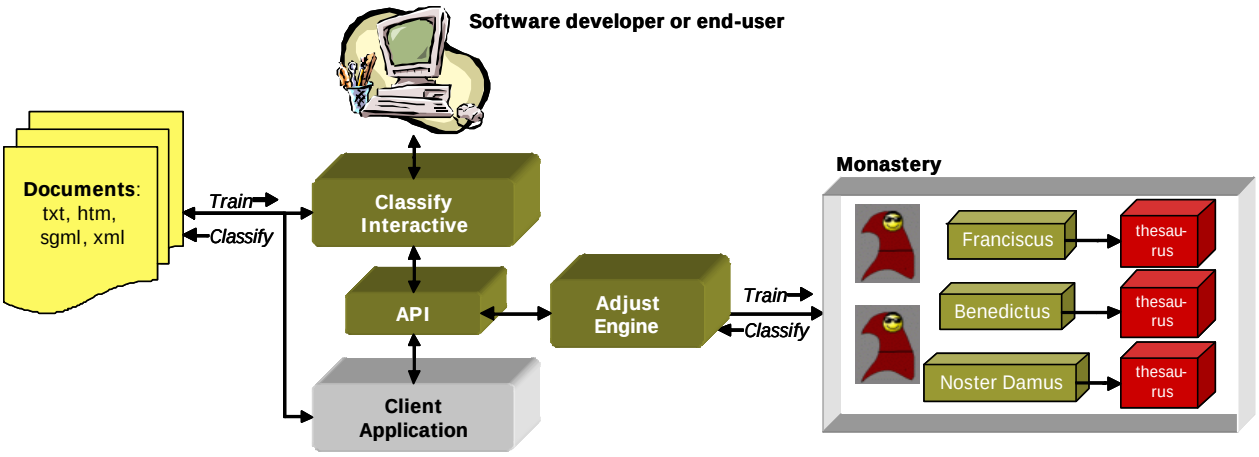

Figure 3: Deploying different classification systems or Monks in an application environment

Currently, separate Monks need to be trained for classifiers in each language. Furthermore, training can be time-consuming. If it is possible to detect the correct meanings of words in the documents, either at training or during classifying, it is possible to apply a Monk trained in one language to documents in other languages. Furthermore, we expect that the training-time decreases because synonyms and variants can be given directly after disambiguation.

3 Resources

3.1 Wordnet

Wordnet is a semantic network, where concepts are represented by synonymous words. The abstract database model of Wordnet is as follows:

- Concept table: list of concept record numbers with concept data:
    - part of speech
    - definitions or glosses,
    - domain labels,
- Concept relations, triplets that consists of *concept-relation-concept*. Examples of relations are:
    - hyponymy: car –has_hyperonym- vehicle
    - meronymy: car –has_part- engine
    - antonymy: close –antonym- open
    - etc.
- Lexical item table: list of words of a language with pointers to concepts, and sometimes lexical information on the lexical items.

At Irion, we use WordNet version 1.6 (Fellbaum 1998). The lexical items table is only available for English and the Domain information is obtained from WordNet Domains (Magnini & Cavagliá 2000).

The next picture shows an overview of the model:

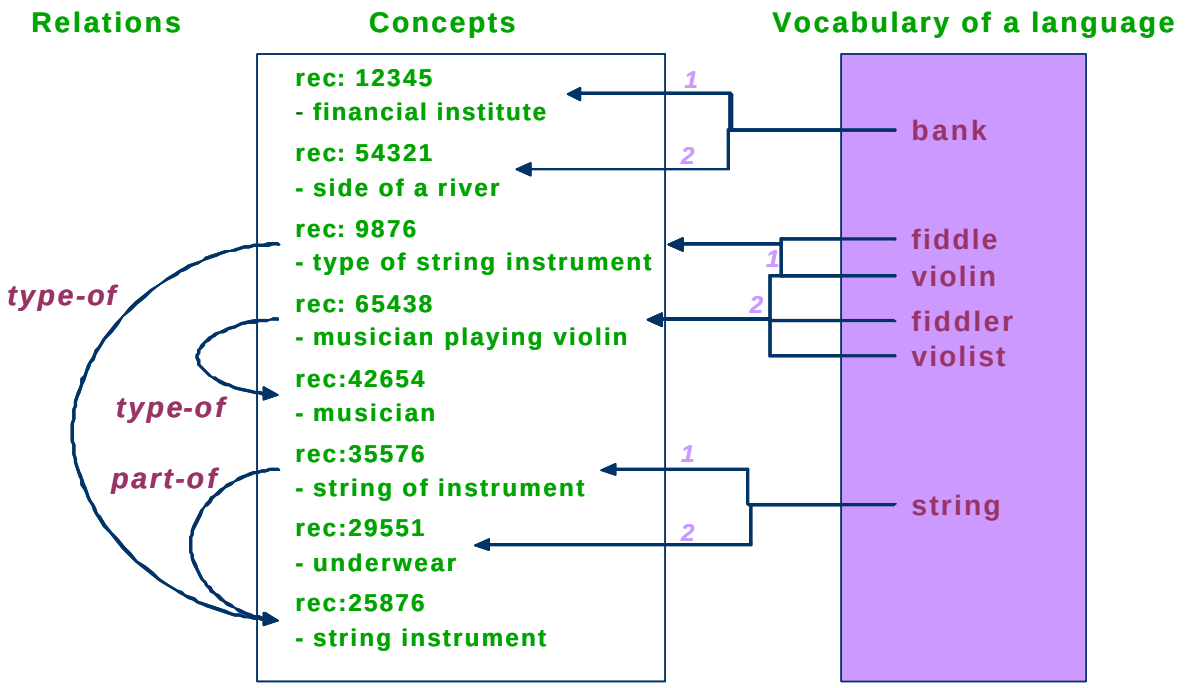

Figure 3: Abstract model of Wordnet

3.2 SemNet

SemNet is a multilingual database that has the same abstract model as Wordnet but has multiple word tables that point to the list of concepts.

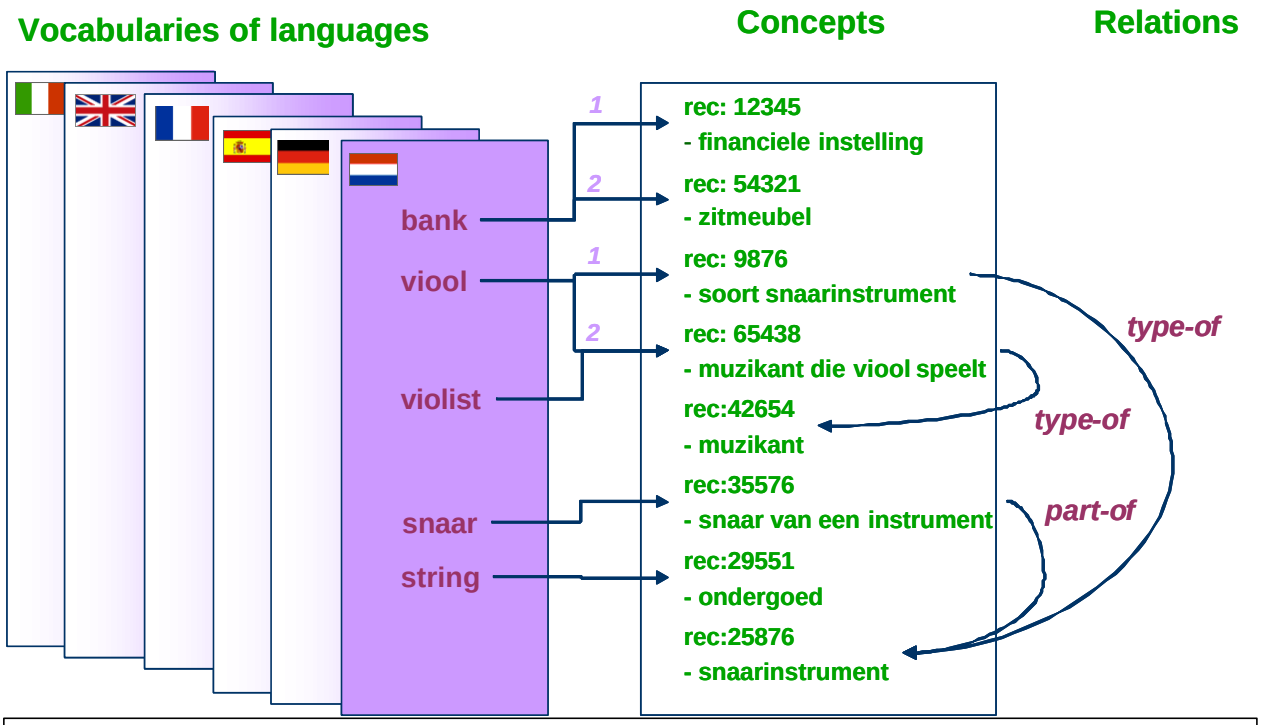

Figure 4: Abstract model of SemNet

SemNet is derived from a multilingual database built by Van Dale publishers. This database goes back to the same database from which the Dutch wordnet was derived. It has a set of Dutch concepts, with Dutch definitions and conceptual relations between the concepts.

The next table gives an overview of the words and concepts in the current version of SemNet:

Table 1: Overview of size of SemNet

	Concepts	Lemmas	Word forms
English	108.154	237.774	410.074
German	101.348	191.210	812.533
Dutch	175.417	213.589	407.360
French	104.373	206.993	433.742
Italian	54.946	81.294	220.482
Spanish	83.059	138.661	255.167
Swedish	36.372	67.972	310.576
Total	175.476	1.137.493	2.849.934

In total, 41,000 concepts are shared by all 6 languages (excluding Swedish).

To compare WordNet with SemNet, the next two tables show the distribution of words, concepts and word meanings for the different Parts Of Speech for WordNet1.6 and the English part of SemNet.

Table 2: Concepts and English lemmas in WordNet1.6

WordNet1.6 English	Concepts	Lemmas	Word meanings	Polysemy	Synonymy
nouns	66025	95135	116364	1.22	1.44
verbs	12127	10326	22074	2.14	0.85
adjectives	17915	19301	28316	1.47	1.08
adverbs	3575	4548	5679	1.25	1.27
Total	99642	129310	172432	1.33	1.30

Table 3: Concepts and English lemmas in SemNet

SemNet English	Concepts	Lemmas	Word meanings	Polysemy	Synonymy
nouns	72459	96461	153458	1.59	1.33
verbs	19551	31756	54767	1.72	1.62
adjectives	9357	17303	35830	2.07	1.85
adverbs	308	580	856	1.48	1.88
Total	101675	146100	244911	1.68	1.44

The resources are globally compatible. The number of concepts and lemmas and the distribution over the different POS are not very different. There are a few striking differences:

- There are hardly any adverbs in the English SemNet;
- The polysemy and synonymy figures of verbs and adjectives are different;

Differences can partly be explained by the fact that the English SemNet is derived from a database based on Dutch words and Dutch concepts. The English translations of these concepts have been used to derive the English SemNet. In Dutch, adverbial used adjectives have the same form and are often not distinguished as a separate part of speech. Consequently the number of adverbs is much lower. Furthermore, more common English words tend to be used more frequently as translation which leads to an increase of the polysemy and synonymy figures.

Still, the overall figures show that it is worth trying to create a mapping between the resources.

3.3 Domains

Both WordNet and SemNet include domain information. The domains for WordNet are taken from WordNet domains (Magnini & Cavagliá 2000). The following domain information is provided for both resources, where we give the number of concepts with a domain label per POS and the proportion compared to all the concepts per POS.

Table 4: Overview of domain information in WordNet and SemNet

	Nouns	Domain Nouns	%	Verbs	Domain Verbs	%	Adjectives	Domain Adjectives	%	Adverbs	Domain Adverbs	%
Wordnet Domains	66025	66025	100%	12127	12127	100%	17915	17915	100%	3575	3575	100%
SemNet Domains	106364	61089	57%	18268	7174	39.%	14514	5998	41%	576	419	72%

We see that WordNet Domains covers all concept whereas the SemNet domains only cover around 50% of all the concepts. The absolute number of Domain labelled concepts are however compatible.

The next table shows the domains from WordNet Domains with a distribution of the number of concepts that are associated.

Table 5: Domain distribution in WordNet1.6

Wordnet Domains	Concepts	Proportion	Wordnet Domains	Concepts	Proportion
acoustics	104	0.092%	linguistics	1545	1.363%
administration	2974	2.624%	literature	686	0.605%
aeronautic	154	0.136%	mathematics	575	0.507%
agriculture	306	0.270%	mechanics	532	0.469%
alimentation	28	0.025%	medicine	2690	2.374%
anatomy	2705	2.387%	merchant_navy	485	0.428%
anthropology	896	0.791%	meteorology	231	0.204%
applied_science	28	0.025%	metrology	1409	1.243%
archaeology	68	0.060%	military	1490	1.315%
archery	5	0.004%	money	624	0.551%
architecture	255	0.225%	mountaineering	28	0.025%
art	420	0.371%	music	985	0.869%
artisanship	148	0.131%	mythology	314	0.277%
astrology	17	0.015%	**number**	220	0.194%
astronautics	29	0.026%	numismatics	43	0.038%
astronomy	376	0.332%	occultism	52	0.046%
athletics	22	0.019%	oceanography	10	0.009%
atomic_physic	66	0.058%	optics	186	0.164%
auto	84	0.074%	painting	123	0.109%
badminton	8	0.007%	paleontology	3	0.003%
banking	102	0.090%	pedagogy	229	0.202%
baseball	144	0.127%	person	1432	1.264%
basketball	44	0.039%	pharmacy	477	0.421%
betting	37	0.033%	philately	6	0.005%
biology	965	0.852%	philology	63	0.056%
body_care	184	0.162%	philosophy	308	0.272%
book_keeping	28	0.025%	photography	124	0.109%
botany	8578	7.570%	physics	1270	1.121%
bowling	33	0.029%	physiology	1096	0.967%
boxing	52	0.046%	plastic_arts	12	0.011%
building_industry	2100	1.853%	play	451	0.398%
card	133	0.117%	politics	1009	0.890%
chemistry	2376	2.097%	post	59	0.052%
chess	28	0.025%	psychoanalysis	67	0.059%
cinema	35	0.031%	psychology	3265	2.881%
color	207	0.183%	publishing	531	0.469%
commerce	636	0.561%	pure_science	96	0.085%
computer_science	481	0.424%	quality	3853	3.400%
cricket	20	0.018%	racing	63	0.056%
cycling	10	0.009%	radio	43	0.038%
dance	143	0.126%	radiology	25	0.022%
dentistry	23	0.020%	railway	15	0.013%
diplomacy	23	0.020%	religion	1598	1.410%
diving	17	0.015%	rowing	8	0.007%

doctrines	73	0.064%	rugby	6	0.005%
drawing	93	0.082%	school	253	0.223%
earth	55	0.049%	sculpture	41	0.036%
ecology	20	0.018%	sexuality	271	0.239%
economy	1437	1.268%	showjumping	1	0.001%
electricity	490	0.432%	skating	18	0.016%
electronics	8	0.007%	skiing	29	0.026%
electrotechnics	37	0.033%	soccer	16	0.014%
engineering	44	0.039%	social_science	1	0.001%
enterprise	372	0.328%	sociology	678	0.598%
entomology	603	0.532%	sport	648	0.572%
ethnology	37	0.033%	state	6	0.005%
exchange	263	0.232%	statistics	3	0.003%
factotum	29015	25.605%	sub	1	0.001%
fashion	936	0.826%	surgery	28	0.025%
fencing	8	0.007%	swimming	53	0.047%
fishing	61	0.054%	table_tennis	22	0.019%
folklore	29	0.026%	tax	101	0.089%
football	90	0.079%	telecommunication	245	0.216%
free_time	360	0.318%	telegraphy	12	0.011%
furniture	545	0.481%	telephony	43	0.038%
gas	5	0.004%	tennis	44	0.039%
gastronomy	2942	2.596%	textiles	1	0.001%
genetics	23	0.020%	theatre	187	0.165%
geography	3099	2.735%	theology	18	0.016%
geology	1143	1.009%	time_period	684	0.604%
geometry	187	0.165%	topography	5	0.004%
golf	84	0.074%	tourism	510	0.450%
grammar	153	0.135%	town_planning	466	0.411%
heraldry	169	0.149%	transport	1704	1.504%
history	1022	0.902%	tv	55	0.049%
hockey	22	0.019%	university	129	0.114%
hunting	153	0.135%	veterinary	1	0.001%
hydraulics	76	0.067%	volleyball	4	0.004%
industry	1101	0.972%	wrestling	32	0.028%
insurance	111	0.098%	zoology	7195	6.349%
jewellery	115	0.101%	zootechnics	63	0.056%
law	1339	1.182%			

In the case of SemNet, the domains have not been assigned for the purpose of word-sense-disambiguation but just for lexicographic purposes. There is a considerable wild-growth of labels and the assignments are not carefully balanced. Some of the subdivision did not seem very useful for selecting meanings either. We therefore grouped the 276 labels into 48 so-called Microworlds. Below, you see the distribution of the Microworlds over the concepts in SemNet.

Table 6: Microworld distribution in SemNet

Microworlds	Concepts	Proportion	Microworlds	Concepts	Proportion
Antrophology	72	0.083%	Hunting	242	0.280%
Archeology	23	0.027%	Industry	2342	2.711%
Architecture	2839	3.286%	Institutes	13	0.015%
Art and Culture	6637	7.683%	Legal	3897	4.511%
Astrology	0	0.000%	Library	52	0.060%
Astronomy	424	0.491%	Linguistics	1717	1.988%
Bio	8451	9.783%	Logic	48	0.056%
Chemistry	1378	1.595%	Materials	1654	1.915%
Clocks	35	0.041%	Mathematics	1266	1.465%
Comestibles	3366	3.896%	Medical	4675	5.412%
Commerce	2595	3.004%	Metereology	577	0.668%
Communication	2178	2.521%	Military	1950	2.257%
Education	1828	2.116%	Nobility	30	0.035%
Environment	220	0.255%	Physics	1374	1.591%
Farming	1759	2.036%	Psychology	2070	2.396%
Finance	3613	4.182%	Religion	2503	2.897%
Fishing	285	0.330%	Sex	132	0.153%
Games	953	1.103%	Society	4175	4.833%
Geography	1180	1.366%	Sport	2997	3.469%
Geology	971	1.124%	Statistics	113	0.131%
Government	2020	2.338%	Technology	4910	5.684%
Grooming	1551	1.795%	Topography	1197	1.386%
History	691	0.800%	Tourists	414	0.479%
House keeping	619	0.717%	Transport	4351	5.037%

You can clearly see that the Microworlds are distinguished at a more global level than the domains in Wordnet. For linking WordNet to SemNet, we made a mapping of the SemNet domain labels with the WordNet domains. This was done at the more detailed label level of SemNet and not at the Microworld level.

Both the WordNet domains and the Microworlds have been used to train a classifier, using TwentyOne Classify Interactive. We extracted all the synonyms of all the concepts and their glosses or definitions that are associated with a domain. These then represent a training document for that particular domain. The same was done for the Microworlds in SemNet. We thus created two monks:

1. Wordius EN: English classifier that associates WordNet Domains with synsets and WordNet glosses;
2. Semius EN: English classifier that associates SemNet Microworlds with English synsets

In Figure-6 below, you see a schematic presentation of this process. The classifiers are based on the words associated with the labels. In addition, we keep an association between the labels and the concepts. Using these classifiers, it is possible to assign a domain or microworld label to an unseen document. Next, we extract the concepts associated with the microworld or domain and select all concepts in the text that match the concepts of the returned microworlds or domains. In Figure-6, a document is classified by Semius as “Sport” and by Wordius as “sport”. We then intersect the concepts of the words in the unseen text with the concepts associated with the microworld and the domain, respectively.

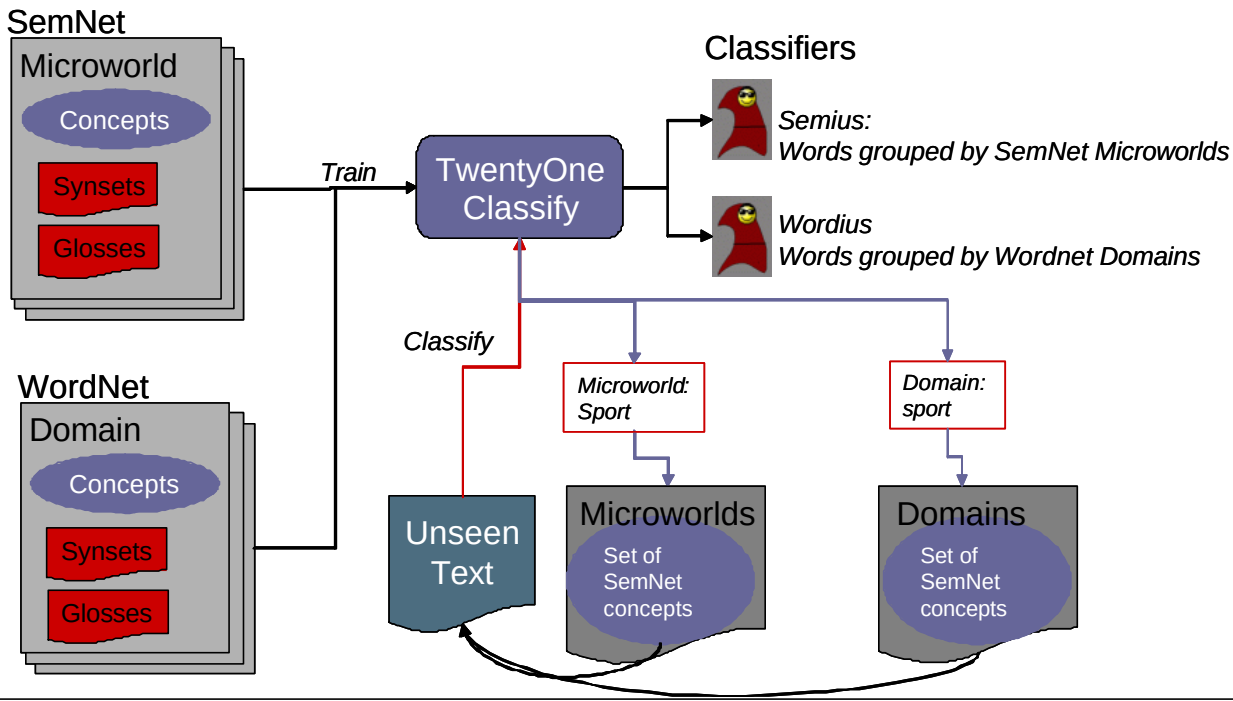

Figure 5: Domain based classifiers derived from SemNet and WordNet

Within this set up, it is easy to switch between WordNet and SemNet by simply replacing the file that associates labels with concepts. In Figure-6, we replaced the concepts of WordNet related to WordNet domains by the concepts of SemNet. Likewise, we can use the classifier derived from WordNet to assign WordNet domain labels to documents but still use the concepts related to SemNet to do the expansion in a particular language. The same can be done the other way around.

The association of SemNet concepts with WordNet domains is based on the automatic mapping that is created between SemNet and WordNet. This is explained in detail in the next section.

Note that SemNet does not have English but Dutch glosses, which can only be used for a Dutch classifier. The English version of Semius is only built using the entries associated with the domain labels. Semius has been used to link WordNet to SemNet and for disambiguating the test collection. Wordius has only been used in the test collection. Using SemNet, it is also possible to make monks for any other language that is associated.

Finally, many of the concepts in WordNet and SemNet belong to the so-called factotum domain (Magnini & Cavagliá 2000). The factotum domain is completely orthogonal with the other domains. Factotum concepts occur in almost every document. In WordNet the label factotum is used explicitly for 25% of the concepts.[1] The same holds for some of the Microworlds in SemNet. The factotum domain and the comparable Mikroworlds (marked in the above tables) have been excluded from the training of the two classifiers.

[1] Domains such as *number* (0.2%) are similar to factocum and will also be orthogonal. However, these domains are very small and have been neglected.

3.4 WordNet linked to SemNet

For MEANING, we linked WordNet 1.6. to SemNet. For this we used the following resources:

- WordNet1.6
- Wordnet Domains
- Dutch wordnet
- SemNet
- Classification system Semius, based on SemNet domain labels grouped into Microworlds

There are two possibilities for matching:

- Dutch lemmas in SemNet to lemmas in the Dutch wordnet
- English lemmas in SemNet to the lemmas in the English Wordnet

Following Atserias et al (1997) and Rigau (1998), we derived a number of heuristics and combined the scores of the individual heuristics per potential link that could be generated by matching the lemmas of WordNet with SemNet.

3.4.1 Heuristics

In total, we applied 13 different heuristics and approaches. Four types of links are extracted from the Dutch WordNet related to the Dutch SemNet and 9 types of links are extracted from the English WordNet related to the English SemNet. The next list shows the identifiers for the links with a short mnemonic name:

01 found_manual_okay
02 found_heuristics
03 unfound_manual_okay
04 unfound_heuristics
05 wn_sn_1synset1sense
06 wn_sn_1sense
07 wn_1sense
08 sn_1sense

09 wn_sn_bttree
10 wn_sn_nttree
11 wn_sn_domain
12 wn_domain_trigger_sn_domain
13 wn_sn_domainlink

Links encoded in the Dutch wordnet (01, 02, 03 and 04)

The links 01/04 were made on the basis of linking results for Dutch nouns and verbs in the Dutch WordNet. All links for the Dutch Wordnet were automatically generated. Many of them were checked manually and found okay. These are called manual_okay. The unchecked ones are called heuristics. They are derived automatically from the available Van Dale lexicons and WordNet1.5. For details on the heuristics see Vossen et al (1997). The links to WordNet1.5 have been converted to WordNet1.6 links, based on the WordNet1.5 to WordNet1.6 mapping in MEANING (Rigau et al 2002).

As we have a newer version of Van Dale than the one that was used for the Dutch Wordnet project, the links were checked in our SemNet database. The versions of the Van Dale database deviated and not all entries and senses could be recovered.

The links from manual_okay that were found in our database with the same VLIS number are stored in found_manual_okay (link type 01), those that were found with a different VLIS number are stored in unfound_manual_okay (link type 02). The same applies to the links derived by heuristics. Links 03 are based on heuristics and were found, whereas 04 are heuristics links that could not be recovered. A small portion was untraceable in our database and was not used. (appr. 7 %)

Links based on mono-semeous entries (05, 06, 07, 08)

Following Rigau (1998) and Atserias et al. (1997), we extracted links by using entries with only a single meaning and or a singleton synset:

- English lemmas in WordNet and in SemNet that have only one meaning and one synset member (05);
- English lemmas in WordNet and in SemNet that have only one meaning and more than one synset member (06);
- English lemmas in WordNet and in SemNet that have one meaning in WordNet and more than one meaning in SemNet (07);

- English lemmas in WordNet and in SemNet that have one meaning in SemNet and more than one meaning in WordNet (08);

In the latter two cases (07, 08) multiple links are extracted for all the senses.

Links based on hierarchical structure (09, 10)

Link numbers 09 and 10 are based on the idea that links on a higher level may be inherited by items on a lower level in the semantic tree. Both heuristics exploit parallelism in the hierarchies Vossen et al (1997).

Link 09 links all identical words from SemNet and WordNet that are linked somewhere higher up in their semantic tree. This is done on the basis of the hyperonym or broader term (BT) relations. For example, the English word “organ” is linked to “musical instrument” in SemNet and WordNet. The hyperonym “musical instrument” already has a reliable link and therefore we cannot the corresponding senses of “organ”. Link 10 is the same as link 09, except that hyponym or narrow term (NT) relations are used for building up the semantic tree.

Links derived using Microworlds and Domains (11, 12, 13)

Both links 11 and 12 use the classifier Semius that is based on the Microworlds in SemNet. Link 11 goes through the WordNet lemmas. Each word, together with its hyponyms plus all the glosses, is sent to the classification system loaded with Semius. Semius returns a set of Microworlds with a score (see previous section). The Microworlds above a threshold have an association with concepts and words in SemNet. If the WordNet lemma matches a SemNet lemma associated with the Microworld, a link is made between the concepts.

Link 12 is derived in a similar way, except that all the words belonging to a WordNet domain are selected together. From these domain words, we keep the monosemous words. Together these are called the trigger words for that domain. The trigger words are sent to Semius. From the returned Microworlds, we collect all the associated words and concepts. For all the polysemous words from the Domain, we then create a link between concepts if the WordNet lemma matches a SemNet lemma associated with the Microworld.

Link 13 is based on a manually association that we created between the WordNet domains and the SemNet domains. Note that this association was not created for the Microworlds because these are too coarse-grained. For each word for which the WordNet domain corresponds to the SemNet domain, a link is made for the corresponding concepts.

The next table shows an overview of the links with a short description, the language of the lemma (N= Dutch and E = English), the reliability of the link and the number of concepts linked. We will explain the reliability figures in more detail below.

Table 7: Overview of WordNet SemNet link types

ID	L	Description	Reliability	nLinks
01	N	Dutch Wordnet – Dutch SemNet, manual_okay links	88.83%	13.428
02	N	Dutch Wordnet – Dutch SemNet, manual_okay links, not recovered	35.60%	11.823
03	N	Dutch Wordnet – Dutch SemNet, heuristics	75.79%	50.445
04	N	Dutch Wordnet – Dutch SemNet, heuristics, not recovered	51.15%	28.761
05	E	An English word has only 1 meaning and that meaning is realised by a single word both in WordNet and in VLIS.	89.08%	4.642
06	E	An English word has only 1 meaning both in WordNet and in VLIS (but WordNet and/or VLIS had synonyms for this meaning)	89.08%	18.580
07	E	An English word has 1 meaning in WordNet (but not in VLIS)	71.98%	37.162
08	E	An English word has 1 meaning in VLIS (but not in WordNet)	72.00%	9.434
09	E	Identical words from SemNet and WordNet that are linked somewhere higher up in their semantic tree. This is done on the basis of hyperonym relations	91.24%	12.679
10	E	Identical words from SemNet and WordNet that are linked somewhere lower down in their semantic tree. This is done on the basis of hyponym relations	77.97%	25.601
11	E	Each word from Wordnet, together with its hyponyms plus all the glosses, is sent to a classifier (which is trained with the SemNet domains). If the found Domain is above the threshold and the word also appears in the list of words of that domain in SemNet, a link is made	48.48%	53.402
12	E	All the words belonging to a Wordnet domain are selected. Words that appear only in 1 synset (have 1 meaning) are called the trigger words for that domain. They are sent to the classifier trained with SemNet domains. If the domain is above the threshold and the word also appears in the list of words of the domain in SemNet, a link is made	68.39%	40.037
13	E	Each word for which the Wordnet domain corresponds to the VLIS domain, a link is made. The link is done on the basis of domain correspondences which were checked manually	80.43%	36.553

The next tables show the number of concepts that have been linked in SemNet and in WordNet on the basis of the above strategy.

Table 8: Coverage of linked concepts in SemNet

SemNet	Nouns		Verbs		Adjectives		Adverbs		Total	
not linked	63282	59.50%	9323	51.03%	8693	59.89%	58	10.07%	81356	58.23%
linked	43082	40.50%	8945	48.97%	5821	40.11%	518	89.93%	58366	41.77%
total	106364		18268		14514		576		139722	

Table 9: Coverage of linked concepts in WordNet

WordNet1.6	Nouns		Verbs		Adjectives		Adverbs		Total	
not linked	32369	49.03%	3770	31.09%	10236	57.14%	2879	80.53%	49254	49.43%
linked	33656	50.97%	8357	68.91%	7679	42.86%	696	19.47%	50388	50.57%
total	66025		12127		17915		3575		99642	

We can see that roughly half of both resources is linked and half is not. We expect that most of the non-linked concepts are rather specific, due to differences in coverage in various domains. To verify this statement, we compared the degrees of polysemy, the number of synonyms, the number of multiword expressions and the average word length, both for the linked and unlinked concepts. We assume the following correlations:

1. specific words tends to have less meanings
2. specific words tend to have few synonyms
3. specific words tend to be multi-words
4. specific words tend to be longer

If the assumptions are correct, then we predict that unlinked concepts tend to have the same features as specific words. The correlations for linked and unlinked concepts are shown in the next tables. The figures in the next tables are derived from SemNet, where we looked at the Dutch and the English lemmas and concepts that are linked and not linked.

Table 10: Correlation between linking and English specificity features in SemNet

English	concepts	forms	word meanings	polysemy	synonymy	multiwords	% multiwords	word length
not linked	49254	108525	124570	1.15	2.20	80245	73.94%	16.3
linked	50388	146186	338060	2.31	2.90	50062	34.25%	10.9

Table 11: Correlation between linking and Dutch specifity features in SemNet

Dutch	concepts	forms	word meanings	polysemy	synonymy	multiwords	% multiwords	word length
not linked	81356	171475	183334	1.07	2.11	61368	35.79%	15.0
linked	58366	158256	228239	1.44	2.71	5152	3.26%	10.5

Here we see that the above claim is confirmed both for Dutch and English. Both polysemy and synonymy figures are much lower for not-linked words and the multi-word and word length figures are much higher. The difference in multi-words between English and Dutch is what we expected because Dutch uses compounds instead of multi-word constructions.

This suggests that most of the unlinked concepts are specific words. As far as they have a single meaning, there is no problem for WSD. As far as expansion is concerned, unlinked WordNet concepts cannot be expanded through SemNet and vice versa but they can also be expanded via their hyponymy relations (through parent translations or grand-parent translations, Atserias et al (1997)).

3.4.2 Evaluation of the mapping

In order to determine the value of each strategy, we took random samples of 100 items for each strategy, and had people check them. The links were presented by:

- The matching lemma in Dutch or in English;
- The part of speech;
- The Dutch concept number and the Dutch definition;
- The WordNet concept number and the English gloss;
- [Optional] the Microworld in SemNet and the domain in WordNet;

Below you see an example of the sample data given for heuristics 12:

```
18 washbasin#n#          152337#kom van een ouderwets wasstel
                         103591393#a bathroom or lavatory sink that is permanently installed
                         and supplied with water
                         Architecture:84#building_industry
19 gas#n#                86626#mengsel van brandstof en lucht dat verbrandingsmotoren aandrijft
```

```
                    110528091#a volatile flammable mixture of hydrocarbons (hexane and
                    heptane and octane etc.) derived from petroleum; used mainly as a fuel
                    in internal-combustion engines
                    Chemistry:93#chemistry
20 pink#n#          11475#elk van de planten die tot het geslacht Dianthus behoren,
                    geteeld voor snijbloemen
                    103885630#a light shade of red
                    Bio:79#color;quality
21 wine_merchant#n# 340384#
                    107666744#someone who sells wine
                    Commerce:90#commerce
22 trace#v#         80478#met lijnornamenten beschilderen
                    300494265#follow, discover, or ascertain the course of development of
                    something; "We must follow closely the economic development is Cuba"
                    Art and Culture:74#doctrines;pure_science
```

We asked two people to verify each sample. They could label the relations as follows:

OK = precise match;
ALMOST = there is a difference in the definition but the differences are not incompatible;
MIGHTBE = the information is too vague or complex to really judge;
WRONG = the information clearly shows that the concepts are different;

If there was no information for one or both concepts, no judgement is made and the link is ignored. The next table shows the results per heuristics, where the figures are averaged over the two reviewers.

Table 12: Evaluation results of linking heuristics

ID	Name	TOTAL SCORED	OK	ALMOST	MIGHTBE	WRONG	OK&ALMOST
1	found_manual_okay	94	65.43%	23.40%	6.38%	4.79%	88.83%
2	unfound_manual_okay	95.5	21.99%	13.61%	20.94%	43.46%	35.60%
3	found_heuristics	95	56.32%	19.47%	9.47%	14.74%	75.79%
4	unfound_heuristics	87	39.66%	11.49%	7.47%	41.38%	51.15%
5	wn_sn_1synset1sense	87	47.70%	41.38%	8.62%	2.30%	89.08%
6	wn_sn_1sense	87	67.24%	21.84%	4.60%	6.32%	89.08%
7	wn_1sense	91	55.49%	16.48%	9.34%	18.68%	71.98%
8	sn_1sense	87.5	50.29%	21.71%	5.71%	22.29%	72.00%
9	wn_sn_bttree	97	70.62%	20.62%	6.70%	2.06%	91.24%
10	wn_sn_nttree	88.5	44.63%	33.33%	10.17%	11.86%	77.97%
11	wn_sn_domain	82.5	16.36%	32.12%	19.39%	32.12%	48.48%
12	wn_domain_trigger_sn_domain	87	59.20%	9.20%	4.02%	27.59%	68.39%
13	wn_sn_domainlink	92	58.70%	21.74%	10.33%	9.24%	80.43%

As we could expect, heuristics 2 and 4 are very unreliable. These concepts could not be resolved across the different versions of the Van Dale database. Surprisingly bad are heuristics 11 and 12. In both cases, we used the SemNet Microworld monk Semius. In the case of 11, Semius classified hyponyms and glosses from WordNet. In the case of 12, Semius classified the monosemeous trigger words from WordNet domains. An explanation for the performance of the 11 could be that hyponyms are usually not limited to a single domain. Heuristics 12 performs somewhat better (almost 70%). It could be the case that many monosemeous domain words from WordNet do not occur in SemNet and thus do not contribute to the results.

Most links are unique. Still quite a few links are confirmed by multiple heuristics. The next table shows the number of words linked by one ore more heuristics:

1 heuristic source	2403118
2 heuristic sources	684162
3 heuristic sources	286508
4 heuristic sources	162003
5 heuristic sources	72719
6 heuristic sources	30207
7 heuristic sources	12647
8 heuristic sources	3060
9 heuristic sources	190
10 heuristic sources	2

(none have more sources)

The maximum number of heuristics is 10 for the Dutch equivalents of “violin”. By combining heuristics, it is possible to derive higher degrees of accuracy. We thus derived a single value for each concept-to-concept link based on the combination of heuristics that contribute to the link.

As a result of the linking between SemNet and WordNet, it is now possible to use the results of MEANING in applying WSD. The next section will explain how WSD is integrated in the Irion applications.

4 Integrating WSD in Applications

Our choice of the approach to WSD is based on the following assumptions:

1. WSD has to be fast;
2. WSD should aim at reducing noise rather than selecting a single sense;
3. WSD should be easily tuneable to different data collections and applications;

The idea is to use the MEANING resources and the WSD technology to mainly reduce the expansion without losing recall.

We make a distinction between concept selection and synonym selection. Concept selection is the process where we select the relevant meaning of a word (or exclude the irrelevant meanings), whereas synonym selection is the process where we select the most likely synonyms from the available concepts. WSD as being applied in MEANING is defined as concept selection rather than synonym selection.

When we consider the distribution of words over text it is clear that we can make some obvious choices when to apply WSD and when not. It is a well known fact that most of the word types are rather specific and have a low frequency. All together they still make up a relatively-small proportion of all the word tokens. These words often have one or just a few meanings and they do not represent a problem for WSD. These words also play a crucial role in retrieval and document classification.

To the contrary, a small set of the most-frequent words have most meanings and make up the largest proportion of a document. It may nog be worthwile to disambiguate words with domain-neutral meanings. Disambiguating these words is computational intensive and has a very low chance of success. The different meanings are vague and tend to be based on collocational and syntactic patterns rather than conceptual relations. This distribution is shown in the next figure:

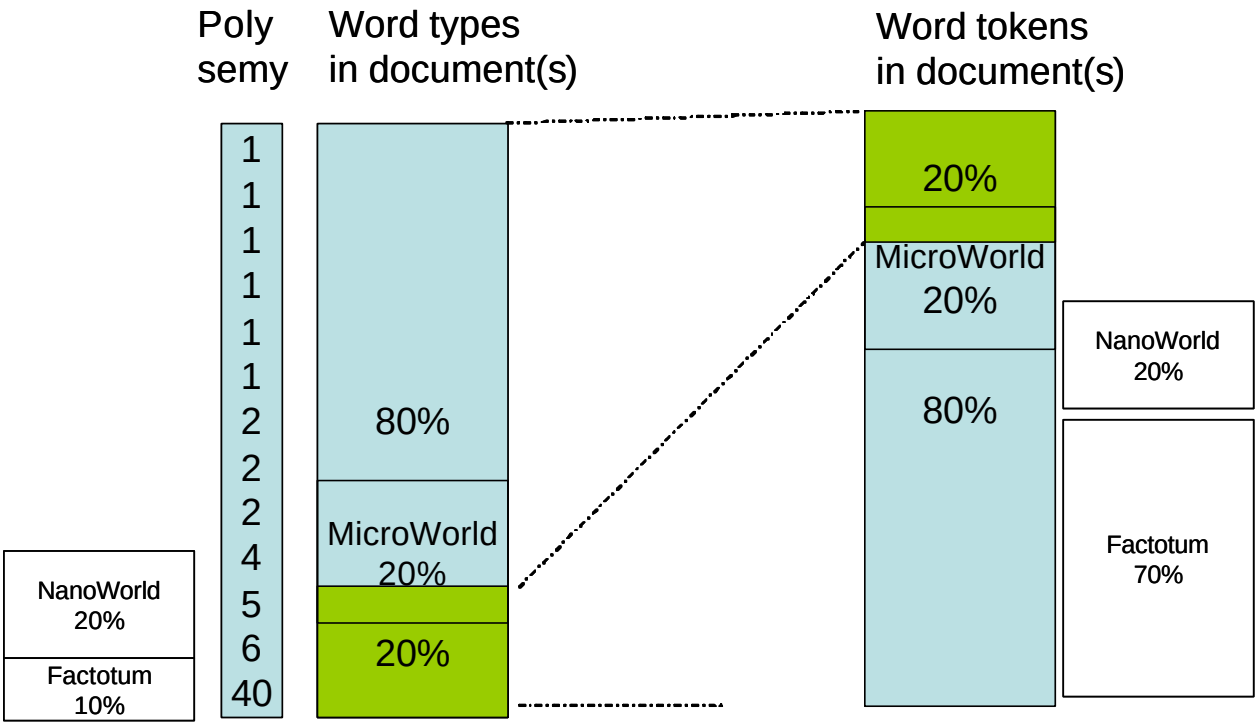

Figure 6 The scope of word-sense-disambiguation

In this figure, you see that many mono-semeous words have low frequencies and a small set of polysemous words have high-frequency. We expect that high-frequent words often are factotum concepts. These are difficult to disambiguate. They occur in complex syntactic structures, have many different meanings and tend to be meaningful only in combination with other words. Many of the low frequent-words will only have one meaning. There will be terminology that cannot be found in the lexicons and many names.

There is then an area at the border line of the most-frequent and low-frequent words where we would like to apply WSD. This area is sensitive to domain-dependencies of meaning and includes words with multiple meanings. Microworlds and domains are expected to make a difference here. When we determine the domains or microworlds that apply to a document or a section of a document, we can select the concepts of words that intersect, as explained above.

In the future (see below), we will try to see if small contexts of the most-frequent words can still be disambiguated using so-called *nanoworld* correlations. Such nanoworld correlations can be derived from selectional-restrictions or non-hyponymic relations that are extracted in

MEANING. For the time being, we treat all frequent and highly-polysemeous words as Factotum words. For these words, we only select the most likely synonyms.

4.1 Current integration of WSD in Irion applications

To integrate WSD in the TwentyOne retrieval application, we apply the following procedure:

- Send the HTML content to the classifier, either Semius or Wordius
- Add the relevant Microworlds or Domains to the META information that goes with each HTM file
- When we extract concepts for indexing:
    - We select only concepts that match the concepts of the associated Microworlds or Domains.
    - We select the most frequent synonyms for polysemeous words that have no relevant domain meaning
    - We select all synonyms for low-polysemeous words

When we built an index, we add expansions related to domains or microworlds or limited by frequency. This works both for the mono-lingual as the cross-lingual indexes.

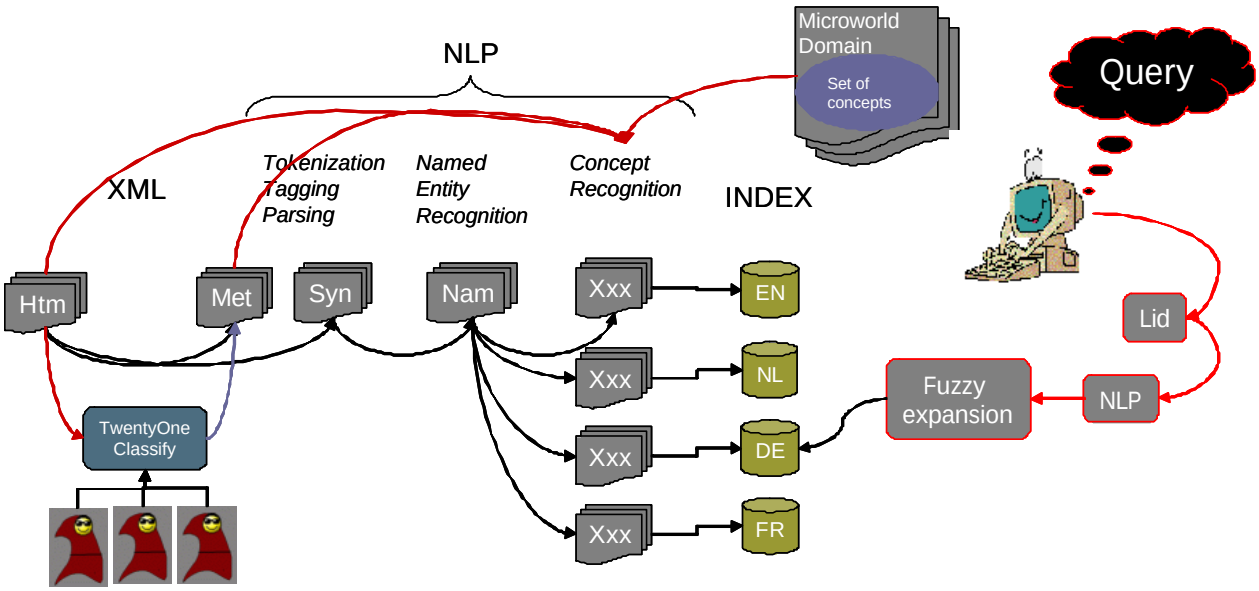

Figure 7: Integrating WSD in Information Retrieval Architecture

For testing, we can tune the expansion in various ways. We can have no expansion, ignore all domains and have full expansion, apply the domains, just apply the frequent synonym selection or combine the latter two.

4.2 Future integration of WSD to Irion applications

There are 4 ways in which we want extend the use of WSD and measure the improvements:

1. Training of domains and microworlds can be extended with topics and examples collected during MEANING;
2. Domains and microworlds can be applied to sequences of NPs in the text. They could represent local maxima of domains that may override the global domain category;
3. Selectional restrictions and non-hyponymy relations extracted in MEANING can be used to expand the concepts associated with the domains and microworlds;
4. Selectional restrictions and non-hyponymy relations extracted in MEANING can be used to derive local maxima in documents that override the global domain;

Training the domains and microworlds makes sense. Currently, the vocabularies associated with the domains and microworlds are very unbalanced. The same domains and microworlds are assigned over and over again, just because these categories are overtrained.

In addition to extending the training set for under-trained domains, we can also ‘purify’ the over-trained domains. This can be done by reducing the vocabulary to the terms that have a high signal value: e.g. long words, monosemeous words, etc..

Options 2 and 4 require the implementation of a different document processing module. First of all, we need to add the domain labels to segments within the text, secondly, we need to change the concept selection procedure to give priority to concepts within the local maxima and only use the global domains if the local maxima are not applicable.

Applying selectional restrictions and non-hyponymy relations to local NPs and VPs may be more complicated. According to Daelemans et al. (2003), training a word expert for small context windows is very time-consuming. It is also unlikely that the classifier can be trained for a huge number of individual concepts related to a few individual concepts. Either the information needs to be generalized to larger sets of concepts with relations or a different process of deciding on correct meanings needs to be designed.

Below, we will discuss the results of the current integration. After the second year of MEANING the above integrations will be implemented and tested with the same data.

5 Test corpus

The Reuters collection contains English news items for the period:

20-August-1996 until 19-August-1997

The files are in XML format. Each file contains text, titles and meta information on the text. For the current tests, we only used the files for the month August 1996. The next table shows an overview of the documents in the test set, which covers the period August, 20-31, 1996.

Table 13: Overview of the Reuters text collection

ID	Doc	HTML Pages	Average Size in KB
1	19960820.xml	2586	1
2	19960821.xml	2521	1
3	19960822.xml	2547	1
4	19960823.xml	2255	1
5	19960824.xml	315	2
6	19960825.xml	495	2
7	19960826.xml	2088	1
8	19960827.xml	2489	2
9	19960828.xml	2643	1
10	19960829.xml	2673	2
11	19960830.xml	2266	2
12	19960831.xml	429	2
Total		23307	1.5

From this collection, we created the following indexes for testing:

1. **No Expansion**: strings are only normalized and literally copied to the other languages;
2. **Full Expansion**: strings are looked up in SemNet and all concepts are selected , all synonyms are expanded (mono-lingually), and all translations are given (cross-linguistically);
3. **Most-frequent synonyms**: no selection of concepts; synonyms are selected on the basis of their frequency as a synonym for all concepts (cut off point is 0.6);
4. **SemNet Mikroworlds**: SemNet Microworlds assigned to documents and concepts selected within the Microworlds if any; all synonyms within Microworld, full expansion if there is no relevant Microworld for a word and it has few meanings;
5. **SemNet Mikroworlds + most-frequent synonyms**: Microworlds assigned to documents and concepts selected with the Microworlds if any; all synonyms within Microworld, if there is no relevant Microworld for a word then we select the most frequent synonym;
6. **WordNet Domains**: same as 4 but with WordNet domains instead of Mikroworlds;
7. **WordNet Domains + most-frequent synonyms**: same as 5 but with WordNet domains instead of Mikroworlds, if there is no relevant Microworld for a word then we select the most frequent synonym;

The next tables show the differences in size for each test set. The databases are differentiated with respect to:

Index:	normalized word to document index
Dictionary:	list of normalized words
Pages:	the XXX files that represent the expanded and translated concepts for the extracted Noun Phrases[2]

The first column gives the numbers for the database with the full expansion. The other tables show the percentages of the other databases relative to the full expansion.

[2] The figures for English pages are slightly higher because the English databases also contain the original HTML files and other intermediate files (SYN and NAM) for extracting the concepts. The databases for the other languages only contain XXX files.

Table 14: Overview of size effects of disambiguation and synonym reduction

		FULL	NO		FRQ		SN_DIS		WN_DIS		SN_DIS_FRQ		WN_DIS_FRQ	
Index (KB)	NL	297705	48228	16.20%	239309	80.38%	190720	64.06%	199291	66.94%	168697	56.67%	178769	60.05%
	EN	401793	46287	11.52%	69814	17.38%	254581	63.36%	268162	66.74%	90268	22.47%	101206	25.19%
	DE	415196	53839	12.97%	232836	56.08%	261102	62.89%	275860	66.44%	193430	46.59%	209131	50.37%
	FR	398281	49618	12.46%	183583	46.09%	256281	64.35%	267701	67.21%	165882	41.65%	179055	44.96%
	IT	239671	44954	18.76%	166120	69.31%	157400	65.67%	164099	68.47%	132816	55.42%	138813	57.92%
	ES	301026	47549	15.80%	170825	56.75%	196010	65.11%	204362	67.89%	147690	49.06%	154532	51.34%
	Total	2053672	290475	14.14%	1062487	51.74%	1316094	64.08%	1379475	67.17%	898783	43.76%	961506	46.82%
Dictionary (KB)	NL	3275	2583	78.87%	3248	99.18%	3245	99.08%	3259	99.51%	3229	98.60%	3240	98.93%
	EN	2505	2077	82.91%	2361	94.25%	2494	99.56%	2499	99.76%	2399	95.77%	2391	95.45%
	DE	3496	2763	79.03%	3410	97.54%	3467	99.17%	3481	99.57%	3414	97.65%	3419	97.80%
	FR	2886	2400	83.16%	2849	98.72%	2875	99.62%	2881	99.83%	2849	98.72%	2851	98.79%
	IT	2926	2430	83.05%	2911	99.49%	2914	99.59%	2920	99.79%	2904	99.25%	2909	99.42%
	ES	2998	2483	82.82%	2968	99.00%	2985	99.57%	2992	99.80%	2966	98.93%	2970	99.07%
	Total	18086	14736	81.48%	17747	98.13%	17980	99.41%	18032	99.70%	17761	98.20%	17780	98.31%
Pages (KB)	NL	968216	141765	14.64%	766901	79.21%	614304	63.45%	646358	66.76%	546271	56.42%	581250	60.03%
	EN	1838307	270595	14.72%	380868	20.72%	1154253	62.79%	1229551	66.88%	470904	25.62%	514353	27.98%
	DE	1461445	160116	10.96%	764915	52.34%	887849	60.75%	948457	64.90%	649612	44.45%	706100	48.32%
	FR	1486099	156728	10.55%	610445	41.08%	911328	61.32%	962422	64.76%	570552	38.39%	618528	41.62%
	IT	784831	133104	16.96%	524156	66.79%	510207	65.01%	536768	68.39%	433516	55.24%	454879	57.96%
	ES	998165	141441	14.17%	534138	53.51%	638496	63.97%	670234	67.15%	478461	47.93%	501007	50.19%
	Total	7537063	1003749	13.32%	3581423	47.52%	4716437	62.58%	4993790	66.26%	3149316	41.78%	3376117	44.79%

If we look at the columns for NO (no expansion and no translation), we see that the size of the index and the size of the pages is about 10 to 15% of the size of the FULL database. This is obvious because the strings are represented as they occur and are only normalized. Note that for the other languages, we simply copied the English words. Interesting observation is that the size of the dictionary is not that much smaller. Apparently, the reduction in word types is much less than the reduction in word tokens.

The FRQ database (only most frequent synonyms are selected, no concept selection), we see that the English index and and pages size is must smaller. This is due to the nature of the function to select the most frequent synonym. In the case of the source word itself, it is obvious that the original word occurs as a synonym in all its meanings and hardly any other word also occurs so frequently in all the meanings. Consequently, the most-frequent synonym function usually only selects the original word. From a cross-lingual point of view, this is different because the polysemy and synonymy are not parallel across languages.

Still, the number of word types in the dictionary is also the same for the English FRQ database. This means that the same words are used but in different combinations and meanings.

A positive effect for WSD is that the index and pages sizes reduced with 40% up to 60%, compared to full expansion. This means 4 times the size of the original data in stead of 10 times. The WSD indexes are also smaller than the FRQ indexes.

6 Cross-lingual Information Retrieval

6.1 Test queries

To develop test queries, Irion first automatically extracts queries from documents. These queries keep track of the database, the document ID and the Noun Phrase on which the query is based. Below is an example of such a query:

```
<TESTIN>
        <DBS_ID>Reuters_2_DIS</DBS_ID>
        <DOC_ID>1</DOC_ID>
        <PAG_TITLE></PAG_TITLE>
        <PAG_ID>2248</PAG_ID>
        <NPS>
        <NP ID="3">detained in a police cell in northern Japan committed suicide by stuffing
toilet</NP>
        </NPS>
        <SOURCE_LNG>en</SOURCE_LNG>
        <BOOLEAN>AND</BOOLEAN>
        <QUERY_EN>detained in a police cell in northern Japan committed suicide by stuffing
toilet</QUERY_EN>
        <QUERY_NL></QUERY_NL>
        <QUERY_DE></QUERY_DE>
        <QUERY_FR></QUERY_FR>
        <QUERY_IT></QUERY_IT>
        <QUERY_ES></QUERY_ES>
</TESTIN>
```

The information tells us that the NP with ID 3 is extracted from page 2248 in document 1 and the database Reuters_2_DIS. Furthermore, the source language of the document is given and the type of query BOOLEAN AND. The latter means that queries will be launched to find all the query words in the document, not necessarily in the same NP. The words from the NP are also used to automatically generate an English query. The assumption is then that this query should give the same page and the same NP as a result. This will be the gold-standard to which we will compare manually derived queries. We assume that it will not be possible to improve the results of launching the ten word NPs as queries to retrieve exactly the same NP. This set is called the “full NP” set.

For the current experiments, we automatically extracted all NPs with 10 or more words. We assume that longer NPs give more possibilities to derive queries. From these queries (more

than 20.000), we selected 100 queries that contain major content words that display a clear ambiguity. The word “cell” in the above query example is such a word.

The 100 queries are edited and translated to extract one and two word queries, which are more realistic. It is important to realize that the absolute results for these queries are rather arbitrary. When we launch a one-word or two-word query, it is not automatically true that the page and the NP of the originally extracted phrase is also the best result. Shorter queries are more sensitive to frequency of the query word in pages and thus may have very different outcomes. In the current system, however, the results are automatically evaluated compared to the original phrases. This is not very fair but less problematic since we only want to compare the systems within the same circumstances.

We extracted the following queries manually:

1. Single word queries:
    a. Ambiguous word, e.g. “cells” in English and a translation in the other languages, e.g. “células”. This is a cross-lingual test compared to the original word in the source language: ***Single word cross-lingual*** (SWX).
    b. A synonym of the ambiguous word, e.g. “neuron” in English: ***Single word synonym*** (SWS).
2. Two word queries:
    a. Ambiguous word combined with a disambiguating context from the NP in English, e.g. “cellular phone”, and translation in the other languages, e.g. “teléfonos celulares”. This is a cross-lingual test compared to the original words in the source language that should be sensitive to WSD: ***Multiword cross-lingual*** (MWX).
    b. Ambiguous word is combined with a synonym of the disambiguating context, e.g. “cellular telephone”; ***Multiword synonym*** (MWS).

In the next example, the queries have been modified manually and translated to other languages.

```
<TESTIN>
        <DBS_ID>Reuters_2_DIS</DBS_ID>
        <DOC_ID>1</DOC_ID>
        <PAG_TITLE></PAG_TITLE>
        <PAG_ID>2248</PAG_ID>
```

```
        <NPS>
        <NP ID="3">detained in a police cell in northern Japan committed suicide by stuffing
toilet</NP>
        </NPS>
        <SOURCE_LNG>en</SOURCE_LNG>
        <BOOLEAN>AND</BOOLEAN>
        <QUERY_EN>police cell</QUERY_EN>
        <QUERY_NL>politiecel</QUERY_NL>
        <QUERY_DE>Polizeizelle</QUERY_DE>
        <QUERY_FR>cellule de police</QUERY_FR>
        <QUERY_IT>cella della polizia</QUERY_IT>
        <QUERY_ES>celda de la policía</QUERY_ES>
</TESTIN>

<TESTIN>
        <DBS_ID>Reuters_2_DIS</DBS_ID>
        <DOC_ID>2</DOC_ID>
        <PAG_TITLE></PAG_TITLE>
        <PAG_ID>743</PAG_ID>
        <NPS>
        <NP ID="22">toxic to nerve cells and activates immune cells in the brain</NP>
        </NPS>
        <SOURCE_LNG>en</SOURCE_LNG>
        <BOOLEAN>AND</BOOLEAN>
        <QUERY_EN>nerve cells</QUERY_EN>
        <QUERY_NL>zenuwcel</QUERY_NL>
        <QUERY_DE>Nervenzelle</QUERY_DE>
        <QUERY_FR>cellule nerveuse</QUERY_FR>
        <QUERY_IT>cellula nervosa</QUERY_IT>
        <QUERY_ES>célula nerviosa</QUERY_ES>
</TESTIN>
```

For an overview of the queries, see the Appendix.

The test system will read a file with the above queries and launch a search. The English query is matched with the English index, the Dutch query with the Dutch index, etc. When the relevant documents have been retrieved for a query, the best 10 documents are selected. We check if the document of the query specification is among the 10 best documents. If so we score this query as a matched document. Secondly, we score each noun phrase of the matched document with respect to the query words. The phrase comparison is based on:

- Fuzziness of the query word matching with the index words
- The number of query words that are matched with a phrase
- The number of phrase words that are not matched with the query word
- Whether or not the phrase words are synonyms or original words

- Whether or not the phrase words are translations or source words

The matching phrases are ranked and we verify if the original phrase word is among the top-three of the ranked NPs. If so, we score this query as a matched phrase. There can thus only be phrase matches within the set of document matches.

6.2 Test results

The next tables show the results of running the queries. The results are differentiated for the above test databases, where we applied no expansion (NO), full expansion (FULL), frequent synonym selection (FRQ), disambiguation (DIS), or a combination of disambiguation and synonym selection (DIS_FRQ). The latter two are differentiated with respect to the classification system that was used for disambiguating. We used the SemNet microworld monk Semius (SN_DIS and SN_DIS_FRQ) and the WordNet domain monk Wordius (WN_DIS and WN_DIS_FRQ).

The first table contains the gold-standard: applying the NPs as queries to retrieve the same NPs. This test only applies to English, because we have a mono-lingual English test corpus. The second column gives the number of unique queries that have been launched. The third column gives the number of correct pages returned within the first 10 pages and and the fifth column the correct number of NPs returned within these pages.

Obviously, we see that the page results for some of these database are close to 100%. This is what we expect for very long NP queries that are extracted from the text itself. There is a slight advantage of the FULL expansion over the other databases.

We also see that the NPs score even better. There is a small flaw here because in some cases there was more than one NP in the same document that matched the query. This explains why all scores are even above the 100%. The differences are marginally.

Table 15: Full NP queries, English

Original NP		EN			
full_np.txt	Queries	Page	%	Np	%
FULL	99	98	0,99	105	1,06
NO	99	96	0,97	103	1,04
FRQ	99	95	0,96	102	1,03
SN_DIS	99	96	0,97	103	1,04
WN_DIS	99	95	0,96	102	1,03
SN_DIS_FRQ	99	95	0,96	102	1,03
WN_DIS_FRQ	99	96	0,97	103	1,04

The second table shows the results for English single word queries, where only the ambiguous word from the original NP is taken (e.g. “cell”). We see a dramatic drop in the results for all the databases (from almost 100% down to 25% up to 40%). Clearly, single word queries can give very different pages compared to the original pages. This is what we expect as well, because there is no reason to prefer the original page over the other returned pages. The frequency of the query word in the document is what counts here and not the context.

Another issue that should be considered here, is that reducing the queries to one word only results in many duplicate queries. The 96 unique NPs only have 38 unique ambiguous head words. Obviously, the retrieval results are the same for each headword but we still measure the correctness by the page and NP from which the original NP was extracted. The same query word “cell” occurring as a head word in a *cell phone* NP and a *police cell* NP will thus get a different scoring when evaluated for one or the other. To minimalize this effect we accept all results within a range of the top-ten results. Still, we estimate that the overall query results drop by about 50% due to this effect. Since this effect is the same for any of the databases that we built, we can still use these queries to measure the differences between the databases.

Table 16: Ambiguous single word query, English

Single word original		EN			
single_word_translation_txt_query.csv	Queries	Page	%	Np	%
FULL	96	25	0,26	19	0,2
NO	96	38	0,4	33	0,34
FRQ	96	31	0,32	24	0,25
SN_DIS	96	27	0,28	23	0,24
WN_DIS	96	24	0,25	18	0,19
SN_DIS_FRQ	96	30	0,31	23	0,24
WN_DIS_FRQ	96	30	0,31	24	0,25

The best results are obtained by not doing any expansion (NO: 40%). This means that all the other approaches introduce more noise than effective expansions. This makes sense since we are looking for the original word that was found in the NP and any expansion will create potential competition.

We do see that the databases with disambiguation improved with respect to the FULL database. Apparently, we are generating less noise and therefore creating less false hits compared to the full expansion. The disambiguated databases did not however perform better than the database without expansion (NO). The gap is about 8-9%. Among the disambiguated databases, the FRQ database performs best and at least as good as FRQ combined with WSD.

Finally, NP matches are equally low as the page matches. This means that within a page match we probably also always have a NP match.

In addition to the ambiguous word, we also created queries by taking a disambiguating context word for English (e.g. “police cell”).

Table 17: Disambiguating multiword query, English

Multiword query		EN			
disamb_cont_np.txt	Queries	Page	%	Np	%
FULL	96	61	0,64	64	0,67
NO	96	76	0,79	77	0,8
FRQ	96	70	0,73	72	0,75
SN_DIS	96	61	0,64	64	0,67
WN_DIS	96	62	0,65	64	0,67

SN_DIS_FRQ	96	68	0,71	70	0,73
WN_DIS_FRQ	96	68	0,71	70	0,73

This shows that more context in the query leads to considerable improvement (around 65% to 79% for page matches). NO expansion scores best but the gap with the disambuated databases got smaller: 6-8%. Furthermore, we see again a clear effect of the disambiguation (6% and 7%) if combined with selection of the most frequent synonym. The latter is not completely fair because the function to extract the most frequent synonym, almost always selects the original word. It generates a full expansion for words with low polysemy and the most frequent synonym for very polysemous words. Apparently, this reduces noise in an effective way, while the full expansion of the low-polysemy words is not very harmful.

The next two tables show the effect of replacing the ambiguous word or replacing the context word by a synonym. If the ambiguous word is replaced by a synonym (cell -> jail), we get the lowest results. There is a hardly no difference across the different databases. Compared to the original single word queries in Table 16, we see that the results drop to 10%. NP matching goes down even more than page matching.

Table 18: Single word synonym, English

Single word synonym (SWS)		EN			
single_word_syn.txt_query.csv	Queries	Page	%	Np	%
FULL	96	10	0,1	4	0,04
NO	96	6	0,06	2	0,02
FRQ	96	10	0,1	5	0,05
SN_DIS	96	11	0,11	4	0,04
WN_DIS	96	10	0,1	2	0,02
SN_DIS_FRQ	96	12	0,12	6	0,06
WN_DIS_FRQ	96	11	0,11	3	0,03

Now we see that the NO expansion has the lowest results. So for synonym queries, the expansion is effective. All disambiguated databases score higher than no expansion.

If we replace the context word by a synonym or related word (e.g. police cell -> prison cell), the results are obviously better. Replacing the context word has a positive effect for the disambiguation (up to 36%), both compared to no expansion (29%) and full expansion (25%). Again the most-frequent synonym selection is most effective.

Table 19: Synonymous context word in multiword query, English

Multi word synonym (MWS)		EN			
disamb_cont_syn.txt	Queries	Page	%	Np	%
FULL	96	24	0,25	22	0,23
NO	96	28	0,29	28	0,29
FRQ	96	35	0,36	33	0,34
SN_DIS	96	25	0,26	23	0,24
WN_DIS	96	23	0,24	22	0,23
SN_DIS_FRQ	96	31	0,32	29	0,3
WN_DIS_FRQ	96	30	0,31	28	0,29

The next tables show the results for cross-lingual queries. Two tests have been done: one where the ambiguous single word query was translated, and the other where the English multiword query with the disambiguating context was translated. The English results given below are for the original English single and multiword queries, as shown in Tables 16 and 17. The English original words are taken as a comparison to measure the effects of cross-lingual search. Note that also for the tranbslated queries there are multiple occurrences of the same query. This effect is however smaller than for English (96:38). The duplications are 96->59 (Dutch), 96->67 (German), 96->53 (French), 96-62 (Italian) and 96->61 (Spanish).

Overall, the results are somewhat better for Dutch, German and French than for Italian and Spanish. This is due to the fact that we did not use a stemmer for the latter languages and the Spanish and Italian semantic networks or wordnets are much smaller.

Also, we see in general that the no expansion database (NO) gives a very poor result for cross-lingual retrieval, whereas it gave best results for mono-lingual retrieval. No expansion for cross-lingual retrieval means that the English source string is put in the index of the other languages as it is. These results are thus obvious.

If we look at the single word queries, we see that the results for Dutch, German and French are lower than for English but not so dramatic. The results drop from 30-40% to around 20% on average. We do see a slight positive effect for the disambiguation but again the frequent synonym selection seems most effective. The results are only 15% to 20% less than for English.

For the multiword queries, the same conclusions can be made. Cross-lingual results drop from 65%-80% to 30-40%. Frequent synonym selection has most effect and disambiguation is less effective than doing full expansion.

Typically, full expansion is as good as doing disambiguation for both the single word queries and the 2-word queries. This is surprising and different from the monolingual results. Apparently, cross-lingual expansion is less harmful than synonym-expansion. It could be that the expanded variants make less sense in combination. It can also be the case that the cross-lingual aspect, that is the mapping of each language to English, is a decreasing factor that does not apply to full expansion.

To summarize: effects of disambiguation as applied here are minimal or even negative in a cross-lingual setting. Most effective is full expansion (!) and selecting the most-frequent synonym or a combination of disambiguation and most-frequent synonym selection.

Table 20: Single word queries, cross-lingual

Single word cross-lingual		EN				NL					DE					FR					IT					ES				
(SWX)	Q	Page		Np		Q	Page		Np		Q	Page		Np		Q	Page		Np		Q	Page		Np		Q	Page		Np	
FULL	96	25	0,26	19	0,2	91	16	0,18	11	0,12	91	18	0,2	10	0,11	91	19	0,21	10	0,11	90	15	0,17	5	0,06	91	18	0,2	9	0,1
NO	96	38	0,4	33	0,34	91	12	0,13	5	0,05	91	8	0,09	6	0,07	91	13	0,14	11	0,12	90	4	0,04	2	0,02	91	3	0,03	1	0,01
FRQ	96	31	0,32	24	0,25	91	19	0,21	13	0,14	91	21	0,23	14	0,15	91	18	0,2	11	0,12	90	13	0,14	4	0,04	91	19	0,21	11	0,12
SN_DIS	96	27	0,28	23	0,24	91	16	0,18	11	0,12	91	16	0,18	9	0,1	91	16	0,18	10	0,11	90	13	0,14	4	0,04	91	18	0,2	10	0,11
WN_DIS	96	24	0,25	18	0,19	91	17	0,19	12	0,13	91	19	0,21	12	0,13	91	16	0,18	10	0,11	90	13	0,14	5	0,06	91	17	0,19	9	0,1
SN_DIS_FRQ	96	30	0,31	23	0,24	91	17	0,19	11	0,12	91	17	0,19	11	0,12	91	18	0,2	12	0,13	90	15	0,17	5	0,06	91	22	0,24	13	0,14
WN_DIS_FRQ	96	30	0,31	24	0,25	91	18	0,2	13	0,14	91	20	0,22	14	0,15	91	16	0,18	11	0,12	90	15	0,17	5	0,06	91	22	0,24	13	0,14

Table 21: Multiword queries, cross-lingual

Multiword cross-lingual		EN				NL					DE					FR					IT					ES				
(MWX)	Q	Page		Np		Q	Page		Np		Q	Page		Np		Q	Page		Np		Q	Page		Np		Q	Page		Np	
FULL	96	61	0,64	64	0,67	96	35	0,36	31	0,32	96	38	0,4	35	0,36	95	42	0,44	38	0,4	94	20	0,21	13	0,14	96	19	0,2	12	0,12
NO	96	76	0,79	77	0,8	96	8	0,08	8	0,08	96	8	0,08	8	0,08	95	10	0,11	10	0,11	94	4	0,04	4	0,04	96	4	0,04	4	0,04
FRQ	96	70	0,73	72	0,75	96	38	0,4	35	0,36	96	37	0,39	35	0,36	95	39	0,41	35	0,37	94	20	0,21	13	0,14	96	18	0,19	13	0,14
SN_DIS	96	61	0,64	64	0,67	96	34	0,35	28	0,29	96	30	0,31	24	0,25	95	36	0,38	32	0,34	94	17	0,18	11	0,12	96	15	0,16	9	0,09
WN_DIS	96	62	0,65	64	0,67	96	30	0,31	24	0,25	96	32	0,33	29	0,3	95	42	0,44	39	0,41	94	19	0,2	14	0,15	96	13	0,14	9	0,09
SN_DIS_FRQ	96	68	0,71	70	0,73	96	35	0,36	28	0,29	96	29	0,3	23	0,24	95	38	0,4	34	0,36	94	16	0,17	10	0,11	96	15	0,16	9	0,09
WN_DIS_FRQ	96	68	0,71	70	0,73	96	31	0,32	25	0,26	96	30	0,31	27	0,28	95	39	0,41	36	0,38	94	19	0,2	13	0,14	96	13	0,14	9	0,09

Q = number of queries.

7 Document Classification

The Reuters collection comes with classification codes that are embedded in the XML structure. There are 3 types of codes:

- Country codes
- Industry codes
- Topic codes

For the current classification experiment, we only used the topic codes. There are 125 different topic codes, which can be organized hierarchically. However, we did not consider the hierarchical relations and treated each code separately. Multiple codes can be assigned to a single document. The Appendix gives an overview of the codes and an explanation.

We extracted the codes for all 23307 documents in the test collection. The Classification system has various options for testing and evaluation. One of the options is that a random test set is extracted from a training set. We thus trained the classification system with 22074 files and set aside a test collection of 233 files. We then constructed the following classifiers from the databases that have been built for retrieval as well:

1. HTM: the plain text is used for training without any linguistic processing
2. NO: the text is linguistically processed but concepts are not expanded
3. FULL: linguistically processed and all NPs are fully expanded
4. FRQ: linguistically processed and all NPs are exp with the synonyms up to 60% of the most frequent synonym.
5. SN_DIS_FRQ: linguistically processed and expanded within SemNet Microworlds or most frequent;
6. WN_DIS_FRQ: linguistically processed and expanded within Wordnet Domains or most frequent;

For each classifier, the same test files are excluded from training. The Classification system can automatically load a folder with test files, classify the files and compare the output of the classification with the classes that are given for the test files. This makes it possible to derive recall and precision figures for the classifiers.

RECALL (ρ) is defined as follows:

$$\rho = \alpha / \tau$$

where:

α = is the number of correct classes assigned to a test file;

τ = is the total number of test classes that are associated with a file;

PRECISION (π) is then defined as:

$$\pi = \alpha / (\alpha + \beta)$$

where:

α = is the number of correct classes assigned to a test file;

β = is the number of wrong classes assigned to a test file;

By assigned we mean that TwentyOne Classify automatically returned a class with a score above the threshold. RECALL and PRECISION are averaged by taking the documents for which there are both evaluation classes and there are classification results (see EVALUATED FILES below). This means that RECALL and PRECISION do not include the files for which no results are given above the threshold.

COVERAGE is then used to indicate on how many file TwentyOne Classify is giving a result. So if the coverage is low, while RECALL and PRECISION are good, you know that you can use this monk to classify a small proportion of the input but that the results are of high quality. If the COVERAGE is high or 100%, then the RECALL and PRECISION figures apply to all the files in the test set as well. If the COVERAGE is low, the figures only apply to those files in the test set for which TwentyOne Classify gives output.

EVALUATION FILES is then the number of files in the test set that had an evaluation class associated. Obviously, there can be files in the test set that do not have such a class.

EVALUATED FILES is the intersection of the COVERAGE and EVALUATION FILES. These are the files with test classes associated and for which results are given above the threshold.

For the experiments, we used a threshold setting of 0.7. This is a bit low compared to other classification data we experimented with (usually 0.8 gives good results). The overall results can thus be easily improved by increasing the threshold to 0.8 or 0.85 without losing much recall and coverage. However, for the comparison of the systems, we can stick to the lower threshold, as long as the threshold is fixed for all systems.

The next table then gives the results for the different classifiers.

Table 22: Recall and precision for Classification

	HTM		FULL		NO		FRQ		SN_ DIS_FRQ		WN_DIS_FRQ	
RECALL	131.6	67.8%	175.5	75.6%	138.8	72.3%	189	81.1%	188.2	80.7%	184.6	79.2%
PRECISION	136.6	70.4%	152.9	65.9%	143.4	74.7%	168	72.1%	168.2	72.2%	166.6	71.5%
COVERAGE	194	83.2%	232	99.5%	192	82.4%	233	100%	233	100%	233	100%
EVALUATION FILES	233	100%	233	100%	233	100%	233	100%	233	100%	233	100%
EVALUATED FILES	194	83.2%	232	99.5%	192	82.4%	233	100%	233	100%	233	100%

The results are compatible with the earlier results given for retrieval. We can take the HTM results as the baseline. In that case, no processing has been applied. We see that FULL expansion leads to an increase of recall and a decrease of precision. This is what we would expect. We also see that the coverage increased: there are more files for which there are results above the threshold.

NO expansion leads to a lower recall (-3.3%) than FULL expansion but remarkably a higher precision (+8.8%). Here we see the effect of just using noun phrase extractions and named-entity recognition. Coverage is lower than for FULL expansion.

Finally, best results are obtained for the disambiguated classifiers and the classifiers expanded with most frequent synonyms. Recall is up to 80% and precision is slightly lower than NO expansion. However, coverage is now 100%. Apparently, the frequency and disambiguation expansion lead to results for documents with words that did not occur in the training set. This can be seen as a positive effect, whereas the negative effect is limited. Overall, we see that the disambiguated expansion in combination with frequency selection can lead to an increase of 12% in recall, 12% in coverage and still 2% increase of precision.

Finally, we have not carried out cross-lingual experiments for classification. This can easily be done by training the classifiers with the cross-lingual data generated for retrieval. However, this would also mean that we need to find similar documents as the test files in each of the languages to compare the results. These documents need to be collected manually or the original documents need to be translated.

8 Conclusions

In this document, we described the integration of WSD in the Irion applications, the linking of WordNet1.6 to the Irion resource SemNet, and the first experiments to validate the effects of WSD for these applications. We applied a simple WSD strategy that is based on the domain information in both WordNet and SemNet.

So far WSD only has marginal effects on retrieval. A positive effect is that the size of WSD indexes are smaller than full expansion indexes. WSD also gives slightly better results for mono-lingual retrieval in the case we use synonyms of words as a query. If we use the original words, no expansion gives best results. In all cases, WSD gives better results than full expansion. For cross-lingual retrieval, no expansion gives worst results. Best results are obtained with full expansion and with WSD.

On the other hand, WSD has a clear positive effect for classification. We increased both precision (2%) and recall (12%) compared to approaches that do not apply NLP.

In the next phrase of the project, we plan to apply the richer data of MEANING to the disambiguation process and also apply different disambiguation strategies to see if the disambiguation leads to further improvements. In the case of the latter, we will investigate systems that can apply conceptual constraints to smaller local contexts such as sequences of NPs and VPs. There are 4 ways in which we want extend the use of WSD and measure the improvements:

1. Training of domains and microworlds extended with topics and examples collected during MEANING;
2. Domains and microworlds applied to sequences of NPs in the text. They represent local maxima of domains that override the global domain category;
3. Selectional restrictions and non-hyponymy relations extracted in MEANING will be used to expand the concepts associated with the domains and microworlds;
4. Selectional restrictions and non-hyponymy relations extracted in MEANING will be used to derive local maxima in documents that override the global domain;

The next evaluation is expected to take place at the beginning of the 3rd year. A final evaluation cycle is planned at the end of the 3rd year of MEANING.

9 References

Atserias, J., S. Climent, X. Farreres, G. Rigau and H. Rodríguez. Combining multiple methods for the automatic construction of multilingual wordnets. In Proceedings of RANLP'97, pages 143-149, Bulgaria, 1997.

Daelemans, W., V. Hoste, I. Hendrickx and A. van den Bosch, 2003, Evaluating the results of a memory-based word-expert approach to unrestricted word sense disambiguation. In Proceedings of the 1st MEANING workshop: Word Sense Disambiguation and Lexical Acquisition, 11th and 12th of April, 2003, Miramar Jauregia, Donostia, Basque Country.

Fellbaum, C. (ed), WordNet. An Electronic Lexical Database, The MIT Press 1998.

Magnini, B. and G Cavagliá, Integrating subject field codes into wordnet. In Proceedings of the Second International Conference on Language Resources and Evaluation LREC'2000, Athens, Greece 2000.

Rigau, G. 1998. *Automatic Acquisition of Lexical Knowledge from MRDs.* Ph.D. Thesis Polytechnic University of Catalonia.

Rigau, G., B. Magnini, E. Agirre, P. Vossen and J. Carroll. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of COLING Workshop, Taipei, Taiwan, 2002.

Rigau, G., B. Magnini, E. Agirre, P. Vossen and J. Carroll. MEANING: a Roadmap to Knowledge Technologies,

Vossen, P. (ed) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht, 1998.

Vossen P, L. Bloksma, P. Boersma, Tools and Resources for the Dutch Wordnet, EuroWordNet Deliverable, D021D025AMS*, University of Amsterdam,* 1997.