Validation of MEANING2 : integration in TwentyOne Search and validation on the EFE Fototeca database

Vossen, P.; Rigau, G.; Alegria, I.; Agirre, E.

2005

document version
Publisher's PDF, also known as Version of record

D8.2 Validation of MEANING2

Version 3.0
29/01/2005

meaning

Developing Multilingual Web-scale language Technologies

INFORMATION SOCIETY TECHNOLOGIES

Work Package 8, Validation and demonstration

Project ref.	IST-2001-34460
Project Acronim	MEANING
Project full title	Developing Multilingual web-scale language technologies
Security (distribution level)	Public
Contractual date of delivery	
Actual date of delivery	
Document number	D8.2
Type	Report
WP contributing to the deliverable	WP8
WP task responsible	German Rigau
EC project officer	Evangelia Markidou
Authors	Piek Vossen (Irion), German Rigau (EHU), Iñaki Alegria (EHU), Eneko Agirre (EHU), David Farwell (UPC), Manuel Fuentes (EFE).
Keywords	Human Language Tecnology, Cross-lingual Information Retrieval, Natural Language Processing, WSD, large-scale knowledge bases.
Abstract	This deliverable reports the integration of the MEANING technology in the TwentyOne search engine of Irion and the initial design and validation of the end-user evaluation framework.

Table of contents

1 Introduction 4

2 TwentyOne Search 5

2.1. Conceptual Search 5

2.2. TwentyOne Search Interface 9

2.3. Cross-lingual architecture of TwentyOne Search 13

3 Integration of MEANING in TwentyOne Search 14

3.1. Integrating Wordnets from the MCR 14

3.2. Integrating WSD in the TwentyOne System 16

4 EFE data 30

5 TwentyOne Search Indexes for EFE 32

6 Design of the experiment 33

6.1. The goal of the experiment 33

6.2. The user tasks 34

6.3. Logging the user-behaviour 36

7 Test pilot results 38

7.1. Automatic retrieval of the test query 38

7.2. User experiences 39

8 Conclusions 42

9 References 43

1 Introduction

This document describes the integration of MEANING in the TwentyOne Search engine of Irion Technologies and the testing of the resulting environment on a real scenario of the Spanish publisher EFE. MEANING acquires lexical knowledge from various sources and various languages. This knowledge is stored in the Multilingual Central Repository (MCR). During the MEANING project, the MCR has been enriched in various cycles. The purpose of work package 8 is to demonstrate that the results of MEANING can be successfully integrated in a real application and on real data.

In deliverable 8.1, we carried out an early baseline demonstration using the semantic networks at Irion and WordNet Domains imported from the MCR. For that purpose, we used the Reuters news collection and a set of 100 queries that were manually created. Originally, Reuters was a user in the MEANING project but they could not participate in the end. During the project, the Spanish publisher EFE agreed to take over the role of Reuters. EFE is interested in setting up a database with news pictures (Fototeca) that can be retrieved by journalists and customers. The database contains pictures with short Spanish or English captions.

In this deliverable, we describe first of all the integration of the MCR in the TwentyOne Search engine of Irion, and secondly, the building of the Fototeca database from two months of captions and pictures. The database will be validated with two experiments:

1. An end-user evaluation to find pictures in the database with and without MEANING results;
2. An automatic evaluation of the database, with and without MEANING, based on a set of about 100 queries;

The task has been devised as follows: a number of end-users will try to find pictures giving an assignment, based on specifically designed sceneries. Their user-behaviour is logged and they will be interviewed.

In this deliverable, we will describe the design of the first experiment and the results of a first pilot test, carried out by a single user. This pilot test will validate the design of the end-user evaluation.

The automatic evaluation will be carried out at the end of the project when having all the data delivered and integrated. The results of this evaluation will be described in deliverable 8.3. Deliverable 8.4 will describe the complete end-user evaluation.

The outline of this deliverable is as follows. We will first describe the TwentyOne Search system, the integration of the data from MEANING and the WSD approach (sections 2 and 3). Section 4 describes the EFE data and section 5 the indexes that have been built. Section 6 explains the set up of the pilot test for the end-user evaluation and section 7 the pilot test results.

2 TwentyOne Search

TwentyOne Search is a conceptual search engine that uses a combination of statistical and language-technology techniques. It is a two step system, where first, the relevant documents are collected using state-of-the-art statistical engines, and secondly, the best matching phrases from the relevant documents are collected. The statistical core-engine of TwentyOne Search returns the most relevant documents from large collections, using a standard vector-space weighting. It ensures fast and robust retrieval. The language-technology has two major roles:

1. Maximize the recall of the statistical engine so that any document is found regardless of the wording and regardless of the query;
2. To extract phrases and concepts from documents so that the best matching phrase can be selected from the relevant documents;

The architecture of TwentyOne Search is such that linguistic processing during querying is minimal but still the benefits of conceptual based retrieval are maximized.

2.1. Conceptual Search

The core idea behind TwentyOne Search is that information is expressed in language by linguistic phrases and not by words in isolation. Consider the following queries:

1. “animal party” versus “party animal”;
2. “Internet servers in Java” versus “Internet servers on Java”

Most traditional search engines will not differentiate these queries and give the same results and most likely not the desired result. The fact that the query words are combined in a linguistic phrase with a particular relation is completely neglected by search engines that look at words in isolation. In the case of *Java the island* it is most likely that you will never find any results because the *Java program language* will dominate the other meanings. Isolated words can have many different meanings hence refer to many different concepts. In combination however, they usually have a very specific and restricted meaning. In addition, very different words can have the same meaning and we also want to retrieve phrases in which these synonyms occur: “beast feast” and “feasting beasts”, or “J2EE Web servers” and “Internet computers on the island Java”.

The TwentyOne Search Engine therefore follows an approach where the text in documents is parsed into linguistic phrases and the words within these phrases are decomposed into sets of relevant concepts. The same applies to the queries that users type in. TwentyOne then matches the concepts related to the user queries with the concepts expressed in linguistic phrases in the documents and it will return those documents first with phrases that include most concepts of the query and that have the best matching concepts. This is shown in the schematic representation of Figure 1.

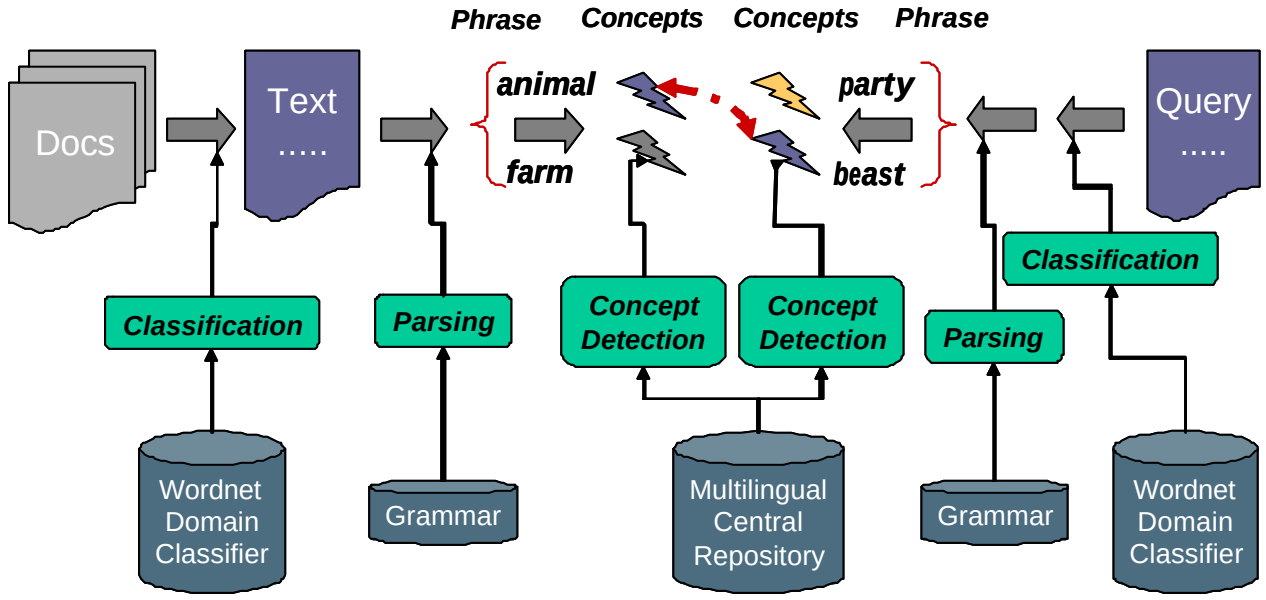

Figure 1: Basic processing scheme for TwentyOne Search

The processing of documents into phrases and concepts roughly takes place in 3 steps:

1. Classification of documents in domains
2. Parsing of the text into linguistic phrases
3. Selection of relevant concepts for the words within a phrase

At the query site, more or less the same process takes place, where classification and parsing are only effective if more than one word is used in the query.

Classification of the domain is done using Irion's text classification system TwentyOne Classify. This is explained in more detail in the next section. Parsing determines in a text what words are used in a linguistic context. The fact that words are used within a constituent structure is highly relevant for determining the actual meaning of the words in context. Within the constituent phrase, we will carry out the selection of the relevant concepts. The concepts are taken from the Wordnet of the relevant language (English and Spanish in the case of the EFE database) and are derived from the MCR. Looking up words includes morphological analysis and stemming.

Concept selection comprises the following analysis:

1. Named entity recognition: *Bush, United States, Pays-Bas*, *The Netherlands*.
2. Multiword lookup: *human rights*
3. Compound resolution: *mensenrechtenactivistenleider* ("mensen-rechten-activisten-leider", literally "human-rights-activist-leader" or leader of the human rights activists)
4. Synonym detection: "beast" and "animal"; "party" and "feast".
5. Word sense disambiguation: "First service ace" {Tennis}; "Early morning service" {Religion}

Named entities are not treated as regular words so that *Bush* is not matched with *plant* and *United States* is also matched with *America, US, USA*, *United States of America, Etats Union* etc. Multiword lookup is important because the combination *human rights* matches with only one concept, whereas *human* and *rights* match with many concepts and concept combinations. Similarly, compounds that do not occur in the Wordnet need to be decomposed into the largest units that do occur. The Dutch example *mensenrechtenactivistenleider* thus includes the English multiword concepts *human right* and *activist leader*.

Finally, the words (single, decomposed or composed units) are related to the concepts using the wordnets in the Multilingual Central Repository (MCR). This means that synonyms are related to the same concept and that relevant meanings are selected if words or multiwords have more than one meaning. Consequently, the system can match occurrences of *animal* with *beast*, and *party* with *feast*, but the latter only when the context is ***not*** politics.

The TwentyOne Search system uses a robust scoring function that matches every query with every phrase. The scoring takes various parameters into account:

1. number of overlapping concepts between the query and the phrase
2. matching domain labels *(optional)*
3. the fuzziness of the query words: “parti” instead of “party”, “aminal” versus “animal”
4. derivational properties of the query word: “feastings”
5. match by synonym or by original
6. part of compound or as isolated word
7. cross-lingual or in the same language

Each of these parameters can be tuned. The scores are combined in a single normalized score. The best matching phrase will correspond with a 100% score and will be similar to an exact search for all the query words, co-occurring in a single phrase. Some small linguistic variation is still allowed for a 100% match (e.g. differences in case, diacritics, hyphens, etc.).

The results are presented by listing the documents with the best matching phrases first. Per document, the phrases that exceed the threshold for the conceptual score are shown. In the case of the EFE interface, the threshold is fixed to 10% as a conceptual score. This means that very poorly matched phrases are also retrieved. The score of the document is however always based on the best matching phrase.[1] It is thus possible to see several phrases per document, where only one phrase represents a high score and the other phrases are of less quality.

[1] In the case of the small EFE articles the number of matching phrases per document is relatively low. There are not many phrases per article that can match the query. It is therefore not necessary to make the threshold tunable in the interface.

The search can be tuned so that it either returns documents with all the query words (boolean AND) or any of the query words (boolean OR). Note that this boolean constraint only applies to the words in a document, using AND does not guarantee that the query words also co-occur in the single phrase. The conceptual phrase matching will remain the same, so that phrases with all the concepts are automatically preferred to phrases with a subset of the concepts. The main effect of using OR instead of AND will be that there are more documents to be considered. This is especially useful when the recall is low for very long or specific queries.

If phrases have the same score, the vector-space weighting of the document (the document-score) is used as a secondary sorting key. Likewise, only the most relevant documents are shown both from a conceptual point of view as from an information value point of view.

Finally, it is possible to use negation operators on words within the query and to combine the queries with searches on meta information that is published, such as date, categories, titles, etc. These features are not further discussed here because they are not relevant for the textual retrieval that is tested in the context of MEANING.

2.2. TwentyOne Search Interface

The TwenyOne Search interface is shown in Figure 2, below. The screen dump shows the result of the query *platos de solomillo troceado*. The query language is Spanish (as opposed to English) and the type of search is the best phrase (la mejor frase), as opposed to exact phrase (la frasa exacta). According to the result list, there are 417 documents that match the query words in total (boolean OR). The maximum number of displayed results is however set to 25 (this can be adapted), so that 25 out of 417 results are presented. The first page lists the first 10 results. Each result consists of:

1. Conceptual score of the match between the best phrase in the document and query, 100% for an ideal match. Note that there can be other phrases in the same document as well that score less than the indicated score;
2. The document identifier, which is the date of the news represented as a single number, e.g. 20040422 = April, 22nd, 2004;
3. The title of the document, here represented by the DESCRIPTION, which is always Spanish, also for English new articles;
4. A flag indicating the source language of the article;

5. The snippets: one or more linguistic phrases with some contexts, where the linguistic phrase is bold and the words from the phrase that match the query are in Italics. All phrases that are above the conceptual threshold setting are shown as a snippet. The score is based on the best matching phrase.

Figure 2: TwentyOne Search interface on the EFE Fototeca database

The screen dump shows that the first result has a phrase match of 100%. The phrase occurs almost literally in the article, except for the difference in case and the difference in plurality. The second document has a lower score of 33% because only 1 out of the 3 concepts in the query occurs in the same phrase.

If a word in the phrase has synonyms or translations, these are shown in a small box when you move the mouse over the italic word. This is shown here for *solomillo* which has as synonyms: ***filete; solomo***.

When we now rephrase the query with the synonym *filete* mistyped as *filette* and leave out the plato, the results look as in Figure 3. The same document is shown first but the score is now lower because the match is based on the synonym. Similary, the second result has a higher score because it contains the *filete.* In both cases the score is a bit lower because of the fuzzy match of *filete* with *filette.*

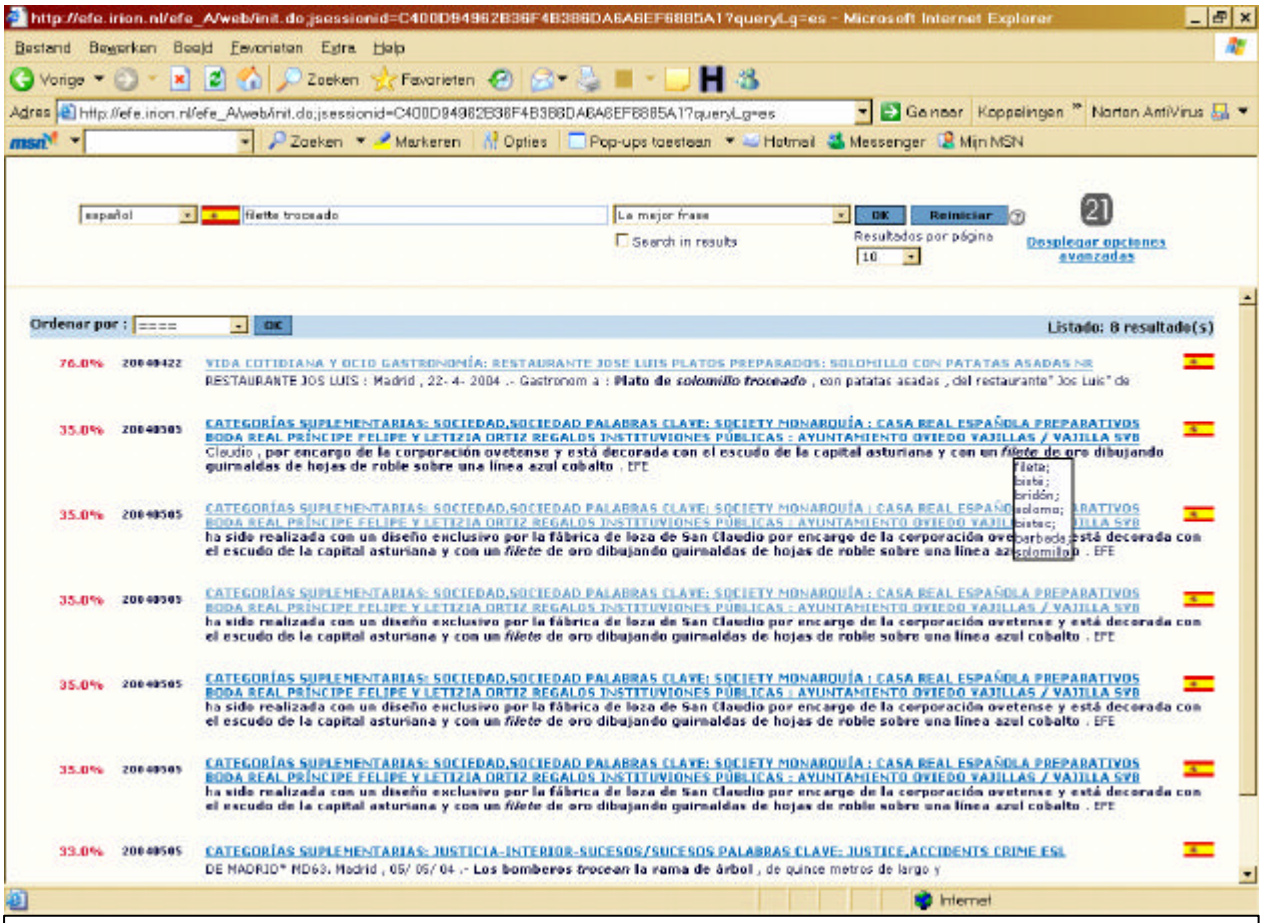

Figure 3: Rephrased query with a synonym

When the user clicks on the blue underlined title of a result, the article and the corresponding picture are displayed in a new window as is shown in Figure 4. The phrases with matching query concepts are highlighted in red. At the bottom of the page, you see three buttons that can be used to express satisfaction, dissatisfaction or uncertainty with the picture that is shown. This information is used to log the user satisfaction with the result.

The upper frame of the page has options to show the meta-information that can be searched for in the advanced mode and to show the original XML file from which the information was derived.

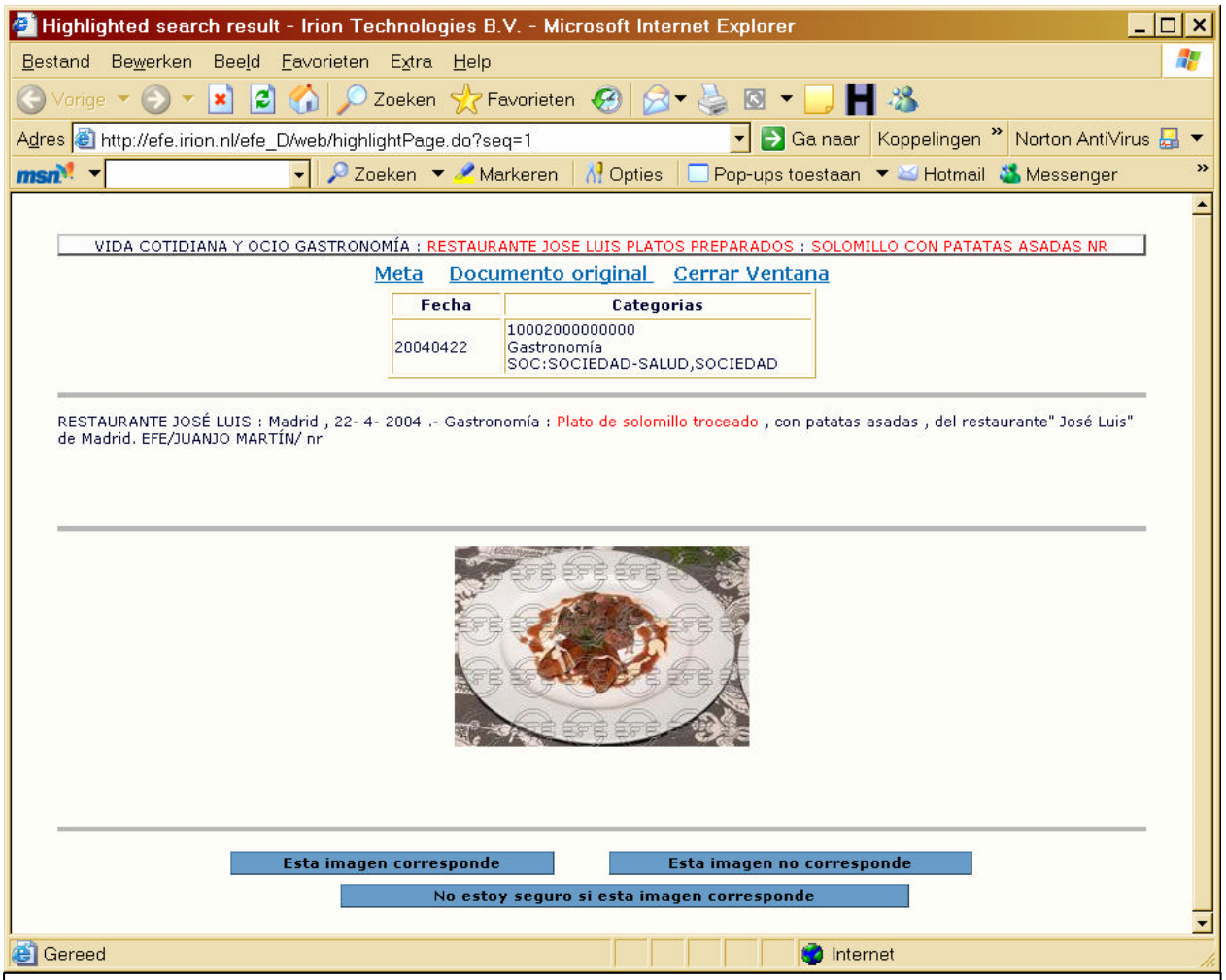

Figure 4: TwentyOne page result with picture from the EFE Fototeca database

2.3. Cross-lingual architecture of TwentyOne Search

For the cross-lingual architecture, we make a distinction between the ***source*** languages and the ***search*** languages. The source languages are the languages of the original documents. In the case of the EFE Fototeca database, these are Spanish and English. The search languages are the languages in which the users can make queries to find documents and phrases from the collection. In the TwentyOne System, the search languages can be different from the source languages, enabling cross-lingual retrieval. This is achieved by building a separate index for every search language, if necessary translating the source language index items..

In the case of the EFE database, we thus first have to process English and Spanish source documents. This includes tokenization, tagging, parsing, named entity recognition and concept extraction. When the concepts are extracted for the English and Spanish source strings, the multilingual semantic network can be used to expand these to English and Spanish synonyms respectively, but also to translations to all the other search languages. Both the source words and the expansions are then normalised and an index is built for each search language based on the normalized words. An overview of this process is shown below in Figure 5.

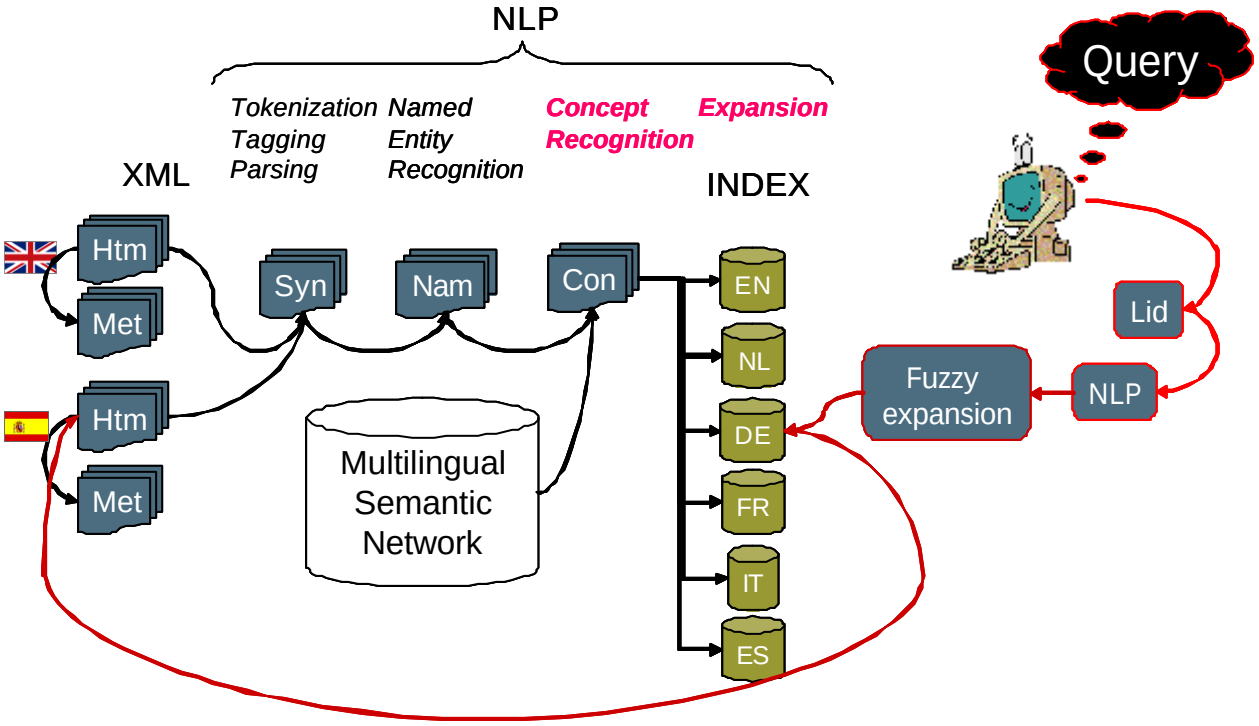

Figure 5: Architecture of TwentyOne Cross-lingual Indexing and Search

During search, the TwentyOne System can detect the query language (or receive it from the interface), next it processes the query words according to the language settings (normalization, compound splitting, etc.) and finally it matches them with the index items of the corresponding language index, in Figure 5, the query is matched with the German index. From the German index it will receive pointers to document identifiers, which correspond with original English or Spanish source files.

3 Integration of MEANING in TwentyOne Search

Integration of MEANING in the TwentyOne Search system involves two different tasks:

1. The use of the wordnets from the MCR to expand the words to synonyms and other variants or related terms;
2. The use of the knowledge in the MCR to build a good word-sense-disambiguation system

3.1. Integrating Wordnets from the MCR

The TwentyOne Search system uses a dedicated database to store the semantic networks and a special program to build up the different indexes for cross-lingual retrieval. The Irion database consists of the following components:

1. A lookup table with word forms and part of speech, linked to a list of concepts;
2. A list of concepts with pointers to synsets;
3. A table with synsets;

Each language has a separate lookup table and a separate table of synsets. All the languages, as much as possible, share the same set of concepts.

To integrate the MCR wordnets, an export was made for the Spanish wordnet consisting of two lists:

1. Lemma + POS + concept number

```
abandonar+v+01524047
```

2. Wordform + POS + concept number

```
abandonara+v+01524047
abandonara+v+01524319
abandonara+v+01525019
abandonara+v+01609431
abandonara+v+01623741
abandonara+v+01728889
abandonara+v+01761339
abandonarais+v+00253929
abandonarais+v+00346044
abandonarais+v+00415168
abandonarais+v+00415444
abandonarais+v+00415625
abandonarais+v+00734233
```

From these lists, a version of the database was built with the Spanish wordnet.

For English, only a lemma list was available from the MCR. The word form table for English was generated by Irion using the word form for the English database that was already available. This could easily be done by matching the lemma+pos information with the existing lemma+pos information and then collect all the wordforms related to the concepts associated with that Irion lemma.

For adding other search languages to the system, we only need a lemma+pos+concept list from the MCR. The wordform lists are only needed when there are also source documents for that language (here only Spanish and English). Likewise, we added the Basque, Italian and Catalan wordnets to the system by importing the lemma+pos+concept exports from the MCR.

3.2. Integrating WSD in the TwentyOne System

Irion adopted the Wordnet-Domains approach to Word Sense Disambigution (Magnini et al 2002), for the following reasons:

1. Easily integrated into the TwentyOne system;
2. High-precision with respect to words that matter for Information Retrieval tasks, e.g. *party* {*free-time* or *politics*};
3. Low-recall with respect to words that do not matter for Information Retrieval tasks, e.g. *part*, *begin*, *be*;
4. Cumulated knowledge from the MCR can easily be added to the system;
5. Fast and robust;

Irion Technology already has a state-of-the-art text classification system: TwentyOne Classify. TwentyOne Classify can be easily trained and benchmarked with any document collection. Likewise, a text classifier can be created by taking any set of words from a wordnet that is associated with a domain. Such a text classifier can then assign domain labels to unseen text, see Figure 6.

For MEANING, we created an option to augment NPs extracted from documents with domain labels. A set of general domain labels is assigned to the document as a whole. This is called a ***microworld*** tag. Next each, NP can get one or more domain labels as well, based on the words from the NP and the words from the surrounding NPs. The NP domains are called ***nanoworld*** tags. It is thus possible that a specific NP in a context has a different set of domains than the document as a whole. You see an overview of this process in Figure 6.

.

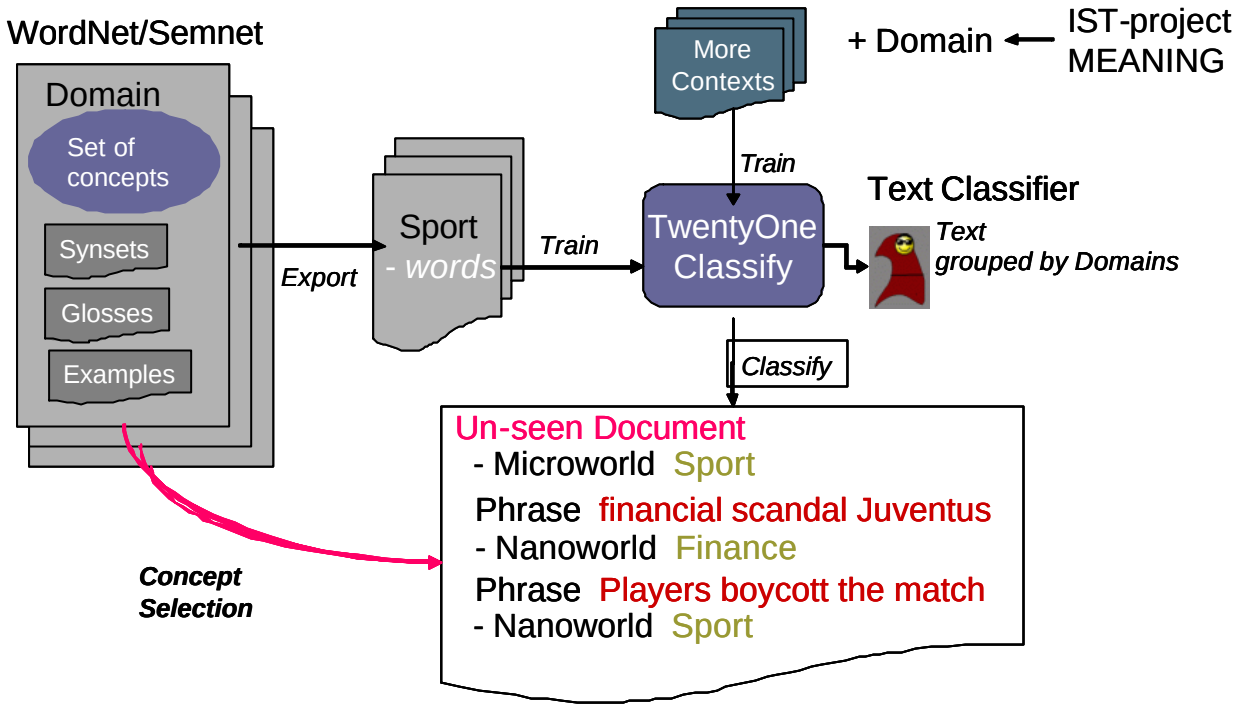

Figure 6: Domain based Word Sense Disambiguation

In the shown example, the document as a whole receives the ***microworld*** tag *Sport* and a specific NP in that document receives the ***nanoworld*** tag *Finance*. The disambiguation then consists of the following process for each word in the NP:

1. Are there word meanings with domain labels that match any of the nanoworld tags? If yes, these meanings are selected.
2. If no, are there there word meanings with domain labels that match the microworld tags? If yes these meanings are selected.
3. If no all meanings are selected.

The text classifiers can be improved in various ways, but within MEANING improvements need to come from the MCR. Cumulation of knowledge in the MCR, from any source in any language, can be easily ported to the classifier by generating word lists that relate this knowledge to domains.

The next page shows a fragment of an English file with NPs tagged with domains. The microworld tags are *art* and *architecture*. A single NP is listed with the id *6*, which in this case, has the same nanoworld tags: *<NW>art;architecture;</NW>*.

```
10\cache_en\151.naw(13):
<MICROWORLD>art;architecture;</MICROWORLD>
<NP ID="6">
<WRD POS="0"><WF>epa00169049</WF></WRD>
<NAME>
<WRD POS="99"><WF>Pete</WF></WRD><WRD POS="99"><WF>Townsend</WF></WRD>
</NAME>
<WRD POS="0"><WF>performs</WF></WRD><WRD POS="99"><WF>on</WF></WRD>
<WRD POS="71"><WF>the</WF></WRD><WRD POS="0"><WF>stage</WF></WRD>
<WRD POS="6"><WF>during</WF></WRD>
<NAME>
<WRD POS="71"><WF>the</WF></WRD><WRD POS="99"><WF>Ronnie</WF></WRD>
<WRD POS="99"><WF>Lane</WF></WRD><WRD POS="99"><WF>Tribute</WF></WRD>
</NAME>
<WRD POS="0"><WF>concert</WF></WRD><WRDPOS="6"><WF>at</WF></WRD>
<NAME>
<WRD POS="71"><WF>The</WF></WRD><WRD POS="99"><WF>Royal</WF></WRD>
<WRD POS="99"><WF>Albert</WF></WRD><WRD POS="99"><WF>Hall</WF></WRD>
<WRD POS="6"><WF>in</WF></WRD><WRD POS="0"><WF>central</WF></WRD>
<WRD POS="99"><WF>London</WF></WRD></NAME>
<PHR>epa00169049 Pete Townsend performs on the stage during the Ronnie Lane Tribute
concert at The Royal Albert Hall in central London</PHR>
<NW>art;architecture;</NW></NP>
```

Separate text classifiers have been made for English and Spanish by exporting the domain vocabulary (based on WordNet Domains version 1.1.1) from the MCR. In total there are 163 domain labels distributed over different levels. Since such a fine-grained domain distinction is not necessary, domain labels at more-specific levels have been merged with the second level domains. For example, the domain *art* has various sublevels that have been lumped together in one domain *art*:

art
- dance
- drawing
  - painting
  - philately
- photography
- music
- plastic_arts
  - sculpture
  - jewellery
  - numismatics
- theatre

We thus created 57 more global domains, as shown in Table 1. The first level domain labels: *applied_science, dictrines, factotum, free_time* and *social_science*, are only represented on so far the are words specifically for these levels.

Table 1: Level 2 Domains

administration	color	mathematics	religion
agriculture	commerce	medicine	sexuality
alimentation	computer_science	metrology	showjumping
anthropology	***doctrines***	military	***social_science***
applied_science	earth	number	sociology
archaeology	economy	pedagogy	sport
architecture	engineering	person	state
art	***factotum***	philosophy	telecommunication
artisanship	fashion	physics	time_period
astrology	***free_time***	play	tourism
astronomy	history	politics	transport
biology	industry	psychology	veterinary
body_care	law	publishing	
building_industry	linguistics	pure_science	
chemistry	literature	quality	

The next two tables show a confusion matrix for these domains, where the words from each domain have been compared with all the other domains by classifying these words. All domains with a score above 60% are listed. We created two tables, one for English and one for Spanish. Overall the confusion matrixes are similar and do not show any striking results. In almost all cases, the domain is associated mostly strongly with itself (scores are mostly 90 or higher). The only exception is *administration*, which has no results above 60 in English and equal scores for *earth* in Spanish. On the second place, we occasionally see other domains with scores of 80 or higher, but then these domains are also strongly related, e.g. *sexuality* and *biology*, or *astronomy* and *astrology*.

It is not very likely that the strongly associated domains will cause any problems in that they will lead to the selection of unintended meanings that will result in bad retrieval. To the contrary, it is probably better to also include the strongly associated domains to have a higher recall. For example, it does not matter if *star* is labelled as *astronomy* or as *astrology* to get the correct interpretation and thus the correct expansion to synonyms or the correct translation, as long as it is not labelled in the *art* or *communication* domain.

This also suggests that the threshold for assigning the domain-labels for Information Retrieval can be relatively low. Whereas normally thresholds of 75 to 85 are required for an optimal text classification it is in this context probably better to work with lower thresholds to have a higher recall and probably harmless overgeneration. The crucial question is to find a balance between recall and precision of the domain assignments so that we find a positive effect for Information Retrieval.

Table 2: Confusion matrix for English domain classifier

earth	6	0,92	earth											
social_science	1	0,75	social_science	0,66	politics	0,64	economy	0,62	doctrines	0,61	sociology			
telecommunication	7	0,91	telecommunication	0,63	architecture	0,61	art							
doctrines	1	0,93	doctrines	0,85	pure_science	0,66	pedagogy							
sexuality	1	0,95	sexuality	0,78	biology									
pure_science	1	0,92	pure_science	0,82	doctrines									
alimentation	2	0,93	alimentation	0,65	biology	0,62	architecture							
medicine	5	0,93	medicine											
sport	30	0,92	sport											
law	1	0,9	law	0,85	economy	0,71	administration							
biology	8	0,94	biology											
industry	1	0,9	industry	0,77	architecture	0,7	economy	0,68	transport					
play	4	0,92	play	0,77	sport	0,7	economy							
tourism	1	0,89	tourism	0,84	transport	0,78	architecture							
number	1	0,94	number	0,67	architecture									
military	1	0,87	military	0,75	history	0,71	transport	0,69	architecture	0,66	earth			
pedagogy	3	0,94	pedagogy	0,62	art	0,6	architecture							
religion	4	0,94	religion											
publishing	1	0,9	publishing	0,77	art	0,71	literature	0,66	economy	0,6	architecture			
administration	1													
metrology	1	0,91	metrology	0,74	physics	0,66	quality	0,64	economy	0,64	time_period	0,62	architecture	
building_industry	1	0,95	building_industry											
artisanship	1	0,92	artisanship	0,77	architecture	0,71	transport	0,65	industry	0,63	mathematics			
astronomy	2	0,93	astronomy											
anthropology	3	0,93	anthropology											
sociology	1	0,92	sociology	0,72	history	0,68	anthropology	0,67	economy					
commerce	1	0,89	commerce	0,89	economy	0,72	architecture							
literature	2	0,93	literature	0,65	publishing	0,63	art							
computer_science	1	0,89	computer_science	0,71	telecommunication	0,66	physics	0,66	engineering	0,6	architecture			
showjumping	1	0,91	showjumping	0,75	sport									
state	1	0,86	state											
body_care	1	0,92	body_care	0,64	architecture	0,61	biology							

physics	7	0,91	physics										
history	2	0,91	history	0,63	anthropology	0,62	sociology						
politics	2	0,92	politics	0,79	economy	0,7	administration						
engineering	5	0,91	engineering	0,7	physics	0,68	transport	0,65	architecture				
factotum	1	0,77	factotum	0,73	architecture	0,71	economy	0,66	transport	0,62	physics	0,61	biology
applied_science	1	0,92	applied_science	0,82	industry	0,66	chemistry	0,62	medicine				
person	1	0,92	person	0,68	economy	0,63	biology						
veterinary	2	0,93	veterinary										
astrology	1	0,93	astrology	0,86	astronomy	0,67	biology						
transport	5	0,92	transport	0,73	architecture	0,63	sport						
chemistry	1	0,89	chemistry	0,75	physics	0,72	biology	0,69	medicine	0,63	architecture		
fashion	1	0,9	fashion	0,69	architecture	0,68	transport	0,66	art	0,64	sport	0,6	biology
economy	8	0,93	economy	0,62	administration								
free_time	1	0,89	free_time	0,86	art	0,67	sport	0,61	pedagogy	0,6	architecture		
archaeology	1	0,93	archaeology	0,76	anthropology	0,72	history	0,62	religion	0,62	biology		
linguistics	2	0,94	linguistics	0,65	economy								
psychology	2	0,93	psychology	0,63	biology	0,6	pedagogy						
agriculture	1	0,91	agriculture	0,88	biology	0,7	medicine	0,69	veterinary	0,66	architecture		
architecture	4	0,92	architecture	0,69	transport								
color	1	0,89	color	0,78	biology	0,71	quality	0,65	earth	0,61	architecture		
mathematics	3	0,93	mathematics										
philosophy	1	0,95	philosophy	0,7	history	0,68	mathematics						
time_period	1	0,92	time_period	0,72	metrology	0,66	economy	0,62	biology				
art	12	0,92	art										
quality	1	0,9	quality	0,72	biology	0,71	metrology	0,63	architecture	0,62	color		

Table 3: Confusion matrix for Spanish domain classifier

earth	6	0,93	earth									
telecommunication	7	0,92	telecommunication	0,61	architecture							
doctrines	1	0,93	doctrines	0,87	pure_science	0,67	pedagogy					
sexuality	1	0,95	sexuality	0,8	biology							
pure_science	1	0,94	pure_science	0,86	doctrines	0,67	pedagogy					
alimentation	2	0,91	alimentation									
medicine	5	0,91	medicine									
law	1	0,93	law	0,83	economy	0,72	administration	0,61	politics	0,61	architecture	
sport	30	0,91	sport									
biology	8	0,94	biology									
industry	1	0,93	industry	0,76	economy	0,74	architecture					
play	4	0,91	play	0,72	sport							
tourism	1	0,91	tourism	0,87	transport	0,79	architecture					
number	1	0,94	number	0,64	architecture							
military	1	0,91	military	0,8	history	0,72	transport	0,72	earth	0,67	administration	0,61 architecture
pedagogy	3	0,95	pedagogy									
religion	4	0,93	religion									
publishing	1	0,93	publishing	0,73	art	0,67	literature	0,6	sport			
administration	1	0,88	administration	0,88	earth	0,8	economy	0,63	law	0,62	architecture	
metrology	1	0,93	metrology	0,73	physics	0,71	quality	0,62	economy	0,6	time_period	
artisanship	1	0,93	artisanship	0,76	architecture	0,72	fashion	0,64	industry	0,62	sport	
astronomy	2	0,94	astronomy									
anthropology	3	0,93	anthropology									
commerce	1	0,92	commerce	0,87	economy	0,69	architecture					
literature	2	0,95	literature									
sociology	1	0,93	sociology	0,7	anthropology	0,7	history	0,65	economy	0,62	politics	
computer_science	1	0,92	computer_science	0,67	physics	0,67	telecommunication	0,62	architecture			
showjumping	1	0,88	showjumping	0,78	sport	0,64	telecommunication					
state	1	0,95	state									
body_care	1	0,94	body_care	0,65	architecture							

physics	7	0,91	physics											
history	2	0,93	history	0,62	earth	0,61	anthropology	0,6	sociology					
politics	2	0,92	politics	0,76	economy	0,7	administration							
engineering	5	0,92	engineering	0,7	physics	0,63	transport	0,61	architecture					
factotum	1	0,84	factotum	0,69	economy	0,67	architecture	0,65	transport	0,63	art	0,63	psychology	
applied_science	1	0,92	applied_science	0,86	industry	0,62	art							
person	1	0,94	person	0,61	economy									
astrology	1	0,9	astrology	0,84	astronomy									
transport	5	0,92	transport	0,63	architecture	0,62	tourism	0,61	engineering					
chemistry	1	0,91	chemistry	0,76	physics	0,68	medicine	0,63	biology					
fashion	1	0,94	fashion	0,69	sport	0,68	transport	0,61	architecture					
economy	8	0,94	economy											
free_time	1	0,92	free_time	0,87	art	0,69	sport							
archaeology	1	0,93	archaeology	0,81	anthropology	0,64	religion	0,63	history					
linguistics	2	0,95	linguistics											
psychology	2	0,93	psychology	0,66	pedagogy									
agriculture	1	0,95	agriculture	0,77	zootechnics									
architecture	4	0,94	architecture											
color	1	0,93	color	0,83	quality	0,71	earth	0,68	biology					
mathematics	3	0,95	mathematics											
philosophy	1	0,95	philosophy	0,79	history	0,68	mathematics							
time_period	1	0,93	time_period	0,65	religion	0,65	metrology	0,6	economy					
art	12	0,93	art											
zootechnics	1	0,92	zootechnics	0,84	agriculture	0,6	economy							
quality	1	0,92	quality	0,72	metrology	0,66	color	0,66	biology					

3.2.1 Effectiveness of Word-Sense-Disambiguation

The classification system used a window of 10 NPs to assign nanoworlds to NPs. This means that 4 NPs to the left and 5 NPs to the right have been used as a context to assign a tag. The microworlds have been assigned to the complete text. For both the nanoworld and microworld tags, we used a threshold of 60. This is relatively low compared to the thresholds that are normally used for text classification (75 to 85). Since the confusion matrix shows that lower level associations of concepts usually also make sense, we think that such a lower threshold is most optimal. Table 4 gives an overview of the number of NPs that received a particular domain tag. The assignment is very distributed. Roughly, we see the same patterns across nanoworlds and microworlds and across English and Spanish. If we take 3% as a threshold, the following domains are consistently above the 3% (or very close): *administration; architecture; art; biology; earth; economy; factotum; history; linguistics; play; politics; religion; sport; time_period; biology; transport*. Some deviation across English and Spanish is found for: *linguistics; pedagogy; and sociology*.

Differences can either be related to differences in the vocabulary associated with WordNet Domain concepts, or to different topics being discussed in the English and Spanish texts. The dominance of some domains (administration, biology, factotum) has also been observed elsewhere (e.g EuroWordNet, Vossen et al. 1998).

Table 4: Wordnet Domain NP tags

	Spanish				English			
Wordnet Domains	Nanoworld		Microworlds		Nanoworlds		Microworlds	
administration	264641	5,09%	15879	5,43%	34440	6,24%	2628	8,46%
agriculture	6376	0,12%	197	0,07%	2618	0,47%	67	0,22%
alimentation	29789	0,57%	900	0,31%	3266	0,59%	63	0,20%
anthropology	144130	2,77%	6994	2,39%	8838	1,60%	410	1,32%
applied_science	234	0,00%	0	0,00%	23	0,00%	0	0,00%
archaeology	2412	0,05%	55	0,02%	362	0,07%	7	0,02%
architecture	270754	5,21%	22221	7,60%	28367	5,14%	2319	7,47%
art	260074	5,00%	18591	6,36%	34071	6,17%	2316	7,46%
artisanship	5917	0,11%	52	0,02%	1323	0,24%	8	0,03%
astrology	642	0,01%	5	0,00%	93	0,02%	2	0,01%
astronomy	34492	0,66%	1366	0,47%	2059	0,37%	15	0,05%
biology	129953	2,50%	7237	2,48%	27921	5,06%	1501	4,83%
body_care	11457	0,22%	126	0,04%	901	0,16%	6	0,02%
chemistry	31428	0,60%	415	0,14%	5447	0,99%	33	0,11%
color	9513	0,18%	409	0,14%	623	0,11%	7	0,02%
commerce	84277	1,62%	3653	1,25%	5944	1,08%	175	0,56%
computer_science	104876	2,02%	2031	0,69%	2850	0,52%	52	0,17%
doctrines	2774	0,05%	181	0,06%	954	0,17%	17	0,05%
earth	260879	5,02%	15215	5,21%	35745	6,47%	2778	8,94%
economy	299046	5,75%	21303	7,29%	33052	5,99%	2419	7,79%
engineering	63163	1,21%	3192	1,09%	5989	1,08%	184	0,59%
factotum	170815	3,29%	9595	3,28%	11644	2,11%	330	1,06%
fashion	64895	1,25%	2645	0,90%	2838	0,51%	62	0,20%
free_time	142776	2,75%	5618	1,92%	11869	2,15%	570	1,84%
history	159952	3,08%	10167	3,48%	18649	3,38%	1154	3,72%
industry	37618	0,72%	1215	0,42%	4343	0,79%	118	0,38%
law	100961	1,94%	5754	1,97%	8308	1,50%	282	0,91%
linguistics	183302	3,53%	14345	4,91%	6136	1,11%	184	0,59%
literature	121272	2,33%	3019	1,03%	7431	1,35%	241	0,78%
mathematics	48858	0,94%	1832	0,63%	4386	0,79%	67	0,22%
medicine	29225	0,56%	795	0,27%	4162	0,75%	64	0,21%
metrology	69650	1,34%	3397	1,16%	15207	2,75%	375	1,21%
military	74417	1,43%	3679	1,26%	15240	2,76%	776	2,50%
number	110936	2,13%	4603	1,57%	10165	1,84%	314	1,01%
pedagogy	87097	1,68%	4670	1,60%	16082	2,91%	1141	3,67%
person	109836	2,11%	5821	1,99%	7280	1,32%	244	0,79%
philosophy	6751	0,13%	59	0,02%	1225	0,22%	10	0,03%
physics	60936	1,17%	2945	1,01%	11979	2,17%	538	1,73%
play	171678	3,30%	9939	3,40%	18560	3,36%	1222	3,93%
politics	213678	4,11%	13803	4,72%	17583	3,18%	1060	3,41%
psychology	49971	0,96%	1244	0,43%	3148	0,57%	39	0,13%
publishing	30297	0,58%	1119	0,38%	5914	1,07%	224	0,72%
pure_science	1890	0,04%	32	0,01%	646	0,12%	4	0,01%
quality	29053	0,56%	741	0,25%	2585	0,47%	13	0,04%
religion	166111	3,19%	9205	3,15%	18793	3,40%	1380	4,44%

sexuality	6903	0,13%	95	0,03%	2645	0,48%	40	0,13%
showjumping	6291	0,12%	133	0,05%	7	0,00%	0	0,00%
social_science	0	0,00%	0	0,00%	62	0,01%	1	0,00%
sociology	163837	3,15%	5033	1,72%	5594	1,01%	128	0,41%
sport	209472	4,03%	14571	4,99%	25463	4,61%	1895	6,10%
state	0	0,00%	0	0,00%	12	0,00%	1	0,00%
telecommunication	151308	2,91%	5642	1,93%	15151	2,74%	879	2,83%
time_period	131129	2,52%	11503	3,94%	14793	2,68%	940	3,03%
tourism	101487	1,95%	6151	2,10%	10976	1,99%	651	2,10%
transport	170083	3,27%	12901	4,41%	17347	3,14%	1066	3,43%
veterinary	0	0,00%	0	0,00%	1002	0,18%	37	0,12%
zootechnics	4611	0,09%	123	0,04%	0	0,00%	0	0,00%
Total	5199312		292293		552111		31057	

The next tables show figures for the concept assignment within the domain tags. More than 2 million word tokens have been looked up in the database. From these, 383,720 word tokens could not be found. More than 1.5 million word tokens were disambiguated, of which 312,699 word tokens effectively (about 20%). Table 5 shows the overall effectivity of the domain tags for these words. For Spanish, about 2,7 million concepts were involved. About 31% of these concepts is not affected by the disambiguation but **69%** is affected. For English, even **82.5%** of the concepts is affected. In general, the recall of the approach is thus very high (!!!). In both cases, the nanoworlds are most effective. This is obvious since we use the microworlds only as a fallback tag in case the nanoworld tags do not apply.

Table 5: Overall effectivity of the domain tags

	Spanish		English	
total concepts	2769753		403124	
disambiguated in microworlds	220574	7,96%	18541	4,60%
disambiguated in nanoworlds	1691079	61,06%	314394	77,99%
unaffected concepts	858100	30,98%	70189	17,41%

In Table 6 and Table 7, we show the reduction and the polysemy. The tables are split over microworlds and nanoworlds. For the microworlds the reduction is about 48% for Spanish and 57% for English. In the case of the nanoworlds, the reduction is even higher: 52% for Spanish and 65% for English.

Table 6: Concept reduction based on Microworlds

Microworlds	Spanish		English	
disambiguated words	44652		3097	
total concepts	220574		18541	
excluded concepts	105620	47,88%	10603	57,19%
selected concepts	114954	52,12%	7938	42,81%
polysemy	4,9		6,0	

Table 7: Concept reduction based on Nanoworlds

Nanoworlds	Spanish		English	
disambiguated words	238671		26279	
total concepts	1691079		314394	
excluded concepts	879317	52,00%	205221	65,28%
selected concepts	811762	48,00%	109173	34,72%
polysemy	7,1		12,0	

The polysemy number of the words affected by the nanoworlds is much higher than the number for the microworlds. This suggests that the more difficult words are solved in a nanoworld context. This is in line with the intuition that the factotum words do not belong to a specific domain and therefore can only be selected in a small context.

The next table also shows the number of times each domain tag was effectively used to select a concept within a microworld and a nanoworld (the 2nd columns). The 3rd columns show the average concept assignment per tag assignment. This indicates how relevant the domain tags have been. We see here that the same tags that are assigned most frequently are also effectively used most frequently. This could suggest that they are simply overasigned and therefore also more effective, in other words: if you shoot often enough you will always hit something. The overassignment is probably due to the unbalanced training of the domain classifier. Some domains have richer vocabularies than others. We cannot conclude from this percentage that the domains are also correctly assigned.

Table 8: Relevance (Rel.) of domain tags per domain

	Spanish						English					
	Nanoworld			Microworld			Nanoworld			Microworld		
Wordnet Domains	Tags	Concepts	Rel.	Tags	Concepts	Rel.	Tags	Concepts	Rel.	Tags	Concepts	Rel.
administration	264641	24004	9%	15879	3583	23%	34440	2038	6%	2628	392	15%
agriculture	6376	246	4%	197	26	13%	2618	146	6%	67	3	4%
alimentation	29789	1637	5%	900	166	18%	3266	345	11%	63	4	6%
anthropology	144130	7770	5%	6994	253	4%	8838	151	2%	410	9	2%
applied_science	234	12	5%	0	0	0%	23	2	9%	0	0	0%
archaeology	2412	28	1%	55	2	4%	362	19	5%	7	1	14%
architecture	270754	22389	8%	22221	2806	13%	28367	3955	14%	2319	209	9%
art	260074	25746	10%	18591	3370	18%	34071	5050	15%	2316	346	15%
artisanship	5917	214	4%	52	0	0%	1323	50	4%	8	0	0%
astrology	642	16	2%	5	0	0%	93	9	10%	2	2	100%
astronomy	34492	611	2%	1366	72	5%	2059	68	3%	15	1	7%
biology	129953	17563	14%	7237	1865	26%	27921	4048	14%	1501	307	20%
body_care	11457	193	2%	126	0	0%	901	42	5%	6	0	0%
chemistry	31428	1879	6%	415	119	29%	5447	244	4%	33	9	27%
color	9513	309	3%	409	10	2%	623	42	7%	7	1	14%
commerce	84277	2369	3%	3653	303	8%	5944	446	8%	175	23	13%
computer_science	104876	1065	1%	2031	130	6%	2850	271	10%	52	5	10%

doctrines	2774	153	6%	181	2	1%	954	18	2%	17	1	6%
earth	260879	20853	8%	15215	2562	17%	35745	2253	6%	2778	245	9%
economy	299046	29130	10%	21303	3249	15%	33052	5261	16%	2419	311	13%
engineering	63163	3849	6%	3192	257	8%	5989	754	13%	184	4	2%
factotum	170815	74987	44%	9595	18556	193%	11644	3435	30%	330	707	214%
fashion	64895	2434	4%	2645	174	7%	2838	226	8%	62	4	6%
free_time	142776	3168	2%	5618	536	10%	11869	408	3%	570	34	6%
history	159952	9747	6%	10167	560	6%	18649	784	4%	1154	54	5%
industry	37618	2291	6%	1215	163	13%	4343	491	11%	118	20	17%
law	100961	8148	8%	5754	879	15%	8308	988	12%	282	87	31%
linguistics	183302	4632	3%	14345	1103	8%	6136	536	9%	184	9	5%
literature	121272	3327	3%	3019	489	16%	7431	417	6%	241	35	15%
mathematics	48858	3024	6%	1832	181	10%	4386	372	8%	67	3	4%
medicine	29225	2054	7%	795	159	20%	4162	412	10%	64	11	17%
metrology	69650	12989	19%	3397	1166	34%	15207	882	6%	375	48	13%
military	74417	9491	13%	3679	695	19%	15240	2011	13%	776	97	13%
number	110936	2032	2%	4603	956	21%	10165	86	1%	314	6	2%
pedagogy	87097	3063	4%	4670	323	7%	16082	1181	7%	1141	75	7%
person	109836	12849	12%	5821	1470	25%	7280	698	10%	244	33	14%
philosophy	6751	299	4%	59	9	15%	1225	27	2%	10	0	0%
physics	60936	3266	5%	2945	183	6%	11979	873	7%	538	24	4%
play	171678	4275	2%	9939	441	4%	18560	2243	12%	1222	55	5%
politics	213678	19729	9%	13803	2295	17%	17583	1909	11%	1060	130	12%
psychology	49971	3067	6%	1244	247	20%	3148	414	13%	39	6	15%
publishing	30297	3017	10%	1119	163	15%	5914	602	10%	224	36	16%
pure_science	1890	103	5%	32	1	3%	646	29	4%	4	1	25%
quality	29053	3046	10%	741	169	23%	2585	376	15%	13	1	8%
religion	166111	5993	4%	9205	962	10%	18793	870	5%	1380	78	6%
sexuality	6903	126	2%	95	20	21%	2645	164	6%	40	2	5%
showjumping	6291	0	0%	133	0	0%	7	0	0%	0	0	0%
social_science	0	0	0%	0	0	0%	62	0	0%	1	0	0%
sociology	163837	6337	4%	5033	1145	23%	5594	427	8%	128	33	26%
sport	209472	25241	12%	14571	1794	12%	25463	6101	24%	1895	265	14%
state	0	0	0%	0	0	0%	12	0	0%	1	0	0%
telecommunication	151308	6242	4%	5642	579	10%	15151	1910	13%	879	41	5%
time_period	131129	22760	17%	11503	2390	21%	14793	1661	11%	940	154	16%
tourism	101487	4896	5%	6151	687	11%	10976	1065	10%	651	71	11%
transport	170083	17530	10%	12901	1698	13%	17347	2877	17%	1066	112	11%
veterinary	0	0	0%	0	0	0%	1002	22	2%	37	0	0%
zootechnics	4611	0	0%	123	0	0%	0	0	0%	0	0	0%
Total	5199312	440199	8%	292293	58968	20%	552111	59709	11%	31057	4105	13%

4 EFE data

EFE delivered an XML file with 29,511 records and 29,943 images. Within each record there is a pointer to an image file. Most of the records contain Spanish articles (26,546) and about 10% is English (2,965). The articles represent two months of news: April – May 2004.

Table 9: Fototeca data from EFE for April-May 2004

		Spanish	English
XML records	29511	26546	2965
Images	29943		

On the next page, an example is given of an English XML records. The actual textual article is found in the field TEXTO. The field IMG_PATH contains the reference to the image:

```
<IMG_PATH>20040406/1152529</IMG_PATH>
```

The initial path directory 20040406 represents the date: 6th April 2004. All pictures of that date are stored in the same subdirectory. The field DESCRIPCION is entered by the EFE editors. It is always Spanish and contains a mixture of keywords and other information. Since there was no specific title field to represent the articles, this field was taken as a provisional title.

A selection of the other fields was chosen as meta information for each file. This means that in the advanced mode, you can use these fields as filters in addition to the textual search. In advanced search mode you can for example search for documents from a specific author <AUTOR>, with a certain orientation <ORIENT> or color <COLOR>, etc. The selection of meta information was made for illustration purposes only. Any other selection can be made.

```
<?xml version='1.0' encoding="UTF-8" ?>

<RECORD>

<ID>FH_1152529</ID>

<CODIGOAUTO>1152529</CODIGOAUTO>

<REVISADO>20040406</REVISADO>

<PAIS_COD>GBR</PAIS_COD>

<CIUDAD>Ross-on-Wye</CIUDAD>

<LUGAR>0</LUGAR>

<AUTOR>David Jones</AUTOR>

<AGENCIA>EPA/PA</AGENCIA>

<FECHA>20040406</FECHA>

<FCREAT>20040406</FCREAT>

<ORIENT>0</ORIENT>

<COLOR>3</COLOR>

<TAMANO>2048x1357</TAMANO>

<PLANO>Plano General</PLANO>

<TEXTO>UK MAZE:epa00167171 High-ho! Off to work goes Edward Heyes as he gives the kilometre of Lawson Cypress of the aMazing Hedge Puzzle its annual trim on Tuesday, 06 April 2004.  Edward and his brother Lindsay built the maze in 1977 at Symonds Yat near Ross-on-Wye in Monmouthshire to commemorate the Silver Jubilee of Queen Elizabeth II.  The two metre high hedges take six weeks to trim - and the sunny side of the hedge gets a second cut later in the year. The maze is arranged in seven concentric octagons and the shortest of 12 routes to the centre is 180 metres, but unwary visitors may discover 13 other routes which are dead ends. Mazes have appeared all over the world since antiquity, the Jubilee Maze is of a traditional design a style known as the 'Labyrinth of Love', which was popular in the heyday of mazes between three and four hundred years ago.  EPA/David Jones UK AND IRELAND OUT[UK AND IRELAND OUT ]</TEXTO>

<TIPO>Temático</TIPO>

<SIGLAS_ENT/>

<DESCRIPCION>CATEGORÍAS SUPLEMENTARIAS: ENTERTAINMENT (GENERAL) LABERINTOS NATURALES / SETOS / ARBUSTOS JARDINERÍA: PODAR EL LABERINTO SE CONSTRUYÓ PARA CONMEMORAR EL JUBILEO DE PLATA DE LA REINA ISABEL II MD**********NO VENDER EN REINO UNIDO**********

**********NO VENDER EN IRLANDA**********</DESCRIPCION>

<CLAS_COD>01016008000000</CLAS_COD>

<IDENTIFICACION/>

<ESTADO>1</ESTADO>

<LOCALIZACION/>

<UBICACION/>

</UBICACION>

<COPYRIGHT>0</COPYRIGHT>

<AMBITO>0</AMBITO>

<ESTILO>Apaisado</ESTILO>

<PERSONAS/>

<COD_AGENCIA>42</COD_AGENCIA>

<SCATE>PA (a partir de 1 mayo 2003)</SCATE>

<CLASIFICACION>Parques, Jardines</CLASIFICACION>

<DESC_ENT/>

<REGIONES>-</REGIONES>

<COMERCIAL>0</COMERCIAL>

<PRODUCT_ID/>

<DREDATE>06/04/2004</DREDATE>
```

```
<HORA>000000</HORA>
<IMG_PATH>20040406/1152529</IMG_PATH>
<TESAURO>CUL:CULTURA-ESPECTACULOS,ARTE SOC:SOCIEDAD-SALUD,SOCIEDAD</TESAURO>
<TIPOAUTO>FOTOTECA</TIPOAUTO>
<IDIOMAAUTO>spanish</IDIOMAAUTO>
</RECORD>
```

5 TwentyOne Search Indexes for EFE

We built 3 different indexes for the EFE Fototeca collection:

- EFE_NO (http://efe.irion.nl/efe_C): no use of wordnets.
- EFE_FULL(http://efe.irion.nl/efe_B): wordnets with full expansion, no disambiguation
- EFE_MEANING (http://efe.irion.nl/efe_A): wordnets with expansion after disambiguation

In the case of the EFE_NO index, the words are not looked-up in the wordnets to expand to synonyms. They are added to the index as they occur in the text, after being normalized according to the language-settings. Note that both the English and the Spanish index thus have the original English and Spanish source words as index items because no translation takes place. Their indexes are thus equal in size and contain exactly the same content.

In the case of EFE_FULL, each word is looked up in the wordnet, where we first check for multiword phrases and apply compound resolution to unknown words. If a word or multiword is found, we take all the meanings and list all the synonyms of that word in addition to the original word. In the case of the cross-lingual indexes, we take all the translations.

In the case of the EFE_MEANING database, we also look up each word and multiword in the wordnet but we try to select meanings within the nanoworld and microworld tags (see above, section 3.2). If these tags apply, we select only the relevant meanings and expand to the synonyms and/or translations of the relevant meanings only. If none of the tags applies, all the word meanings are taken, just as with EFE_FULL.

The size of the indexes thus correlates with the amount of expansion. The EFE_NO index will be the smallest and the EFE_FULL index will be the biggest. We can thus measure the effectiveness of the Domain-based disambiguation by measuring the size of the index and the index vocabulary relative to EFE_NO and EFE_FULL. This is shown in the next table:

Table 10: Indexed lemmas for different EFE indexes

Lemmas	NO	FULL	MEANING
English	53872	70703	69716
Spanish	53872	63352	62644

The full expansion thus enlarged the lemma list with 17,000 for English and almost 10,000 for Spanish. The disambiguation then reduces the list with about 1000 items for both languages.

6 Design of the Pilot Test

6.1. The goal of the experiment

This section describes the initial design of the end-user evaluation framework. The evaluation will be performed in a real scenario provided by Spanish news agency EFE. During the meeting in Madrid we discussed several possible scenarios. In particular, we decided to investigate a multilingual database of pictures: FOTOTECA. This database receives about 800 pictures every day. Each picture has a short caption mainly in Spanish and English. Now, these captions are manually translated for multilingual access.

EFE has provided to us a small sample of two month of text captions and the associated pictures. Some key points about the EFE scenario:

- They receive around 800 pictures every day.
- There are Spanish (from EFE) and English texts (from EPA and AP).
- EFE is translating manualy most of the English texts.
- 50 words per text on average.
- Users usually ask for Named Entities: Persons locations and Events.
- The text is in XML format.

MEANING has designed a complete end-user evaluation framework. This design has been validated in a Pilot Test with a single user from EHU. Mainly, the user was asked to perform

a set of tasks with different systems in a limited time. Finally, the user was asked to fill in a questionnaire.

With this pilot test, we plan to check the appropriateness and correctness of the whole evaluation framework including, the tasks design, the questionnaire, the systems, the logging files, the number of necessary end-users, etc. If necessary, we will modify the end-user evaluation framework for the final test accordingly. The end-user final evaluation test will be reported in Deliverable D8.4.

6.2. The user tasks

The end-user pilot evaluation test has been designed to be performed by a single user. The end-user will test three different systems (namely, efe_no, efe_full, efe_meaning). Each test has four different tasks. That is, the end-user will test a total number of sixteeen different tasks using three different systems. The total time to perform each test has been set to tweenty minutes. After finishing each test, the end-user will have another ten minutes for answering a common questionnaire. That is, a total time of thirty minutes for each test, and one hour and a half as an estimated total time for completing the whole Pilot Test. We include in appendix A the three test sets, and in appendix B the questionnaire.

Each test set has been designed to be self-content described. First, we suggest to the end-user to read the instructions carefully. Then, we inform the end-ser that he is preparing four articles with accompanying pictures and a system located in a particular web page provides him access to the EFE Fototeca database, the system is accessible using Internet Explorer versions 5.0 and higher. We provide to the end-user access to the system using a particular username and password. In particular, he is preparing a news article of TOPIC about CONTEXT, and he feels that the text would be well served at this point by a visual showing GOAL as it is shown in the example News Article 3.

News Article 3
TOPIC = ECONOMIA
CONTEXT = El petróleo está subiendo de precio descontroladamente y hay un debate político internacional sobre las causas y las soluciones.
GOAL = Un político hablando sobre el precio del carburante.

In the task of News Article 3, the end-user is required to locate a picture showing a politician talking about the fuel prices (GOAL), in the context of the international uncontrolled rising of fuel prices (CONTEXT) within ECONOMY (TOPIC).

Now, he should query the Fototeca database using the system we are testing for an appropriate photograph. He should revise the results and select the appropriate picture. When he finds an appropriate photograph, he should click on the button labeled “This is the right picture”, but If this picture is not appropriate, he should click on the button labeled “This is the wrong picture”. We inform the end-user that If he do not find an appropriate photo the first time, he can try modifying the query, adding, removing or changing words from the original query. He can also select more than one picture for each news article. However, the total time for locating the appropriate pictures for the four news articles of one test is only twenty minutes. When finishing the end-user will have another ten minutes for answering a questionnaire.

We have prepared an common end-user questionnaire to be filled after finishing each test set of four tasks. The questionnaire consists of nine closed questions, allowing the end-user to provide at the end additional comments. Each question is in fact an statement that the user can agree or disagree providing a numerical score (1 = strongly agree, 2 = agree, 3 = have no opinion, 4 = disagree, 5 = strongly disagree). The statements where:

1. The instructions were clear.
2. I succeeded in getting what I wanted done.
3. The queries were normal and natural.
4. The system understood what I wanted say.
5. At each point during the interaction I understood what I could say.
6. The system behaved as expected.
7. The interaction was very long.
8. I had particular trouble with queries for:
    a. Types of people,
    b. Types of places,
    c. Types of objects or artifacts,
    d. Types of events or activities,
    e. other
9. I would use this system again to help me find a photo.

Finally we thank the end-user for evaluating the system.

6.3. Logging the user-behaviour

For the experiment, the logging of TwentyOne Search was adapted to derive the results for the experiment. For each of the indexes, a separate log file is created that stores the actions of the users and the results, e.g.:

```
tester1 # Tue Jan 11 11:28:33 CET 2005 # 5E0D57ACC0B87CA52CEAE850E767C9C9 # SEARCH : es : terrorista : 597 : 25 : 5\45 : 17\291 : 28\334 : 6\252 : 46\113 : 26\402 : 7\452 : 14\4 : 26\401 : 26\378

tester1 # Tue Jan 11 11:28:38 CET 2005 # 5E0D57ACC0B87CA52CEAE850E767C9C9 # HIGHLIGHT : es : 5\45 : 20040404/1150676

tester1 # Tue Jan 11 11:28:42 CET 2005 # 5E0D57ACC0B87CA52CEAE850E767C9C9 # UNDECIDED : es : 5\45

tester1 # Tue Jan 11 11:28:45 CET 2005 # 5E0D57ACC0B87CA52CEAE850E767C9C9 # HIGHLIGHT : es : 17\291 : 20040416/1162724

tester2 # Tue Jan 11 11:28:46 CET 2005 # 83BAB57609C13A5BF5EC2083FBCA02D2 # HIGHLIGHT : es : 27\40 : 20040425/1172234

tester1 # Tue Jan 11 11:28:48 CET 2005 # 5E0D57ACC0B87CA52CEAE850E767C9C9 # CONFIRMED : es : 17\291

tester2 # Tue Jan 11 11:28:57 CET 2005 # 6FD00A5A4EA5F328EA3F90BD83AE44B8 # SEARCH : es : Panama : 262 : 25 : 46\298 : 53\354 : 46\297 : 7\205 : 34\139 : 34\138 : 34\360 : 34\361 : 53\353 : 34\140

tester2 # Tue Jan 11 11:29:04 CET 2005 # 6FD00A5A4EA5F328EA3F90BD83AE44B8 # SEARCH : es : Panama barca : 345 : 25 : 5\226 : 5\227 : 8\271 : 5\370 : 42\165 : 9\196 : 8\336 : 13\209 : 56\147 : 56\149

tester2 # Tue Jan 11 11:29:10 CET 2005 # 6FD00A5A4EA5F328EA3F90BD83AE44B8 # HIGHLIGHT : es : 5\226 : 20040404/1151306

tester2 # Tue Jan 11 11:29:12 CET 2005 # 6FD00A5A4EA5F328EA3F90BD83AE44B8 # DISAPPROVED : es : 5\226

tester2 # Tue Jan 11 11:29:16 CET 2005 # 6FD00A5A4EA5F328EA3F90BD83AE44B8 # HIGHLIGHT : es : 5\227 : 20040404/1151307

tester2 # Tue Jan 11 11:29:19 CET 2005 # 6FD00A5A4EA5F328EA3F90BD83AE44B8 # DISAPPROVED : es : 5\227
```

On each line you see a request which has the following syntax:

- name of the person logged in#

- time stamp#

- session id#

- action, where there are the following actions: SEARCH, HIGHLIGHT, DISAPPROVED, UNCERTAIN, APPROVED.

The SEARCH action has the following syntax:

- query language:

- query string:

- number of total results:

- number of collected results:

- top ten results displayed on the first page: document-id\page-id

The number of total result is the total set of articles that contain the query words. In the case of boolean AND, these are the articles with all the query words, in the case of boolean OR the articles with any of the query words. The search engine has a maximum number of results for which it carries out the conceptual match of the query with the NPs. The total result is cut-off by the maximum results. Currently, the maximum is set to 25 but it can be any number. A higher maximum will only slow down the searches. On this set, the conceptual match is carried out and the matches above the threshold represent the collected results.

The HIGHLIGHT ACTION has the following syntax:

- result language:

- result id: document-id\page-id:

- picture reference (if any):

The DISAPROVED/APPROVED/UNCERTAIN actions have the following syntax:

- result language:

- result id: document-id\page-id

The users are instructed to press the new task button after completing each task. The new task button will assign a new session ID to the user. Each session ID thus marks a different

task. In the above example from the log file, there are two different users: tester1 and tester2. We see that one action of tester2 interrupts the sequence of actions of tester1. We also see that tester2 started a new session in the last part of the log file.

7 Pilot test results

7.1. Automatic retrieval of the test query

We used the benchmarking environment of TwentyOne Search to test the effectiveness of the system to find the articles with the correct images from the 21 tasks, taking the correct search field. In total 25 queries are created because some of the 21 tasks had multiple correct results. In that case each result was considered separately. We measured the recall that the correct article was listed among the first ten results. We tested this on all 3 indexes. The results are shown below in Table 11.

Table 11: Retrieval results with pilot queries

Queries	Experiment					
Query file	D:\Irion\TwentyOneTest\EFE_3_queries.xml					
Nr. Queries	25					
Recall	NO		FULL		MEANING	
top 10	19	0,76	20	0,8	20	0,8
1	13	0,52	16	0,64	16	0,64
2	5	0,2	3	0,12	3	0,12
3	1	0,04	1	0,04	1	0,04
4	0	0	0	0	0	0
5	0	0	0	0	0	0

We see here that the use of wordnets is effective. The total recall of EFE_NO is 4% lower and the recall for the first position is even 12% lower. However, we did not measure any difference between EFE_FULL and EFE_MEANING. This means that the disambiguation did not result in a measurable recall effect for these queries.

7.2. User experiences

The main purpose of the Pilot Test was in fact to obtain feedback from a user prespective with respect the current design of the end-user evaluation scenario.

In table 12, we provide the approximated time (in minutes) the end-user spend per task. The total time per system A (efe_no) was more than three times larger than using system B (efe_full). And both the system A (efe_no) and C (efe_meaning) performed similarly. However, when system A (efe_no) and system B (efe_full) have small differences between each of their respective tasks (2 minutes for system A and 1 minute for system B), it seems system C has large variations (7 minutes). This would indicate that task 2 and task 4 for system C (News Article 17 and 21 respectively) had some kind of problematic phenomena.

We should remark that the tasks (News Articles) the end-user was testing for each system were different. That is, task 1 using system A, B or C was different and can not be directly compared.

Table 12: Time in minutes (approx.) per task

	System A (efe_no)	System B (efe_full)	System C (efe_meaning)
Task 1	6	2	2
Task 2	4	1	8
Task 3	4	1	1
Task 4	4	1	6
Total	18	5	17

In fact, in the task designed for News Article 17 the GOAL was “Unos empleados de un centro comercial preparando el puesto de hortalizas” (Workers in a commercial center preparing the vegetables stand), and the end-user focus the queries on the commercial center without success. Regarding News Article 21, the end-user was not sure about the correct picture that was retrieved in a short time because the picture is not clear at all (the pictures have been manipulated with the EFE logo).

Obviously, the problems with particular tasks will be the same for all systems when running the Final Test allowing us to obtain comparable results for the systems.

Table 13 contains the answers of the end-user questionarie. Recall the numerical scores indicate the degree of agreement with each statement (1 = strongly agree, 2 = agree, 3 =

have no opinion, 4 = disagree, 5 = strongly disagree). With respect the design of the test, question 1 seems to indicate that the instructions were clear (except an error with the web address of system B) Question 2, 4, 6, 5, 7 and 9 indicate that the best system for the point of view of the user was system B (efe_full), and the worst system C (efe_meaning) because of the system did not behave as expected (question 6) or the interactions were very long (question 7). Although the user spent more time with system A than with C, the user consider that the interaction was longer using system C than A (question 7). However, using system C the user obtained in some cases the results very fast, but in others after multiple interactions. This problem (which in fact the responsability corresponds to the design of the the News Article task) seems to affect the whole impression of the user with respect system C (questions 2, 4, 5 and 7).

Table 13: Answers for the end-user questionarie

Question	System A	System B	System C
1. The instructions were clear.	1	1	1
2. I succeeded in getting what I wanted done.	2	1	4
3. The queries were normal and natural.	2	2	2
4. The system understood what I wanted say.	2	1	3
5. At each point during the interaction I understood what I could say.	2	1	3
6. The system behaved as expected.	3	1	4
7. The interaction was very long.	4	5	2
8. I had particular trouble with queries for:			
a. Types of people,	5	4	2
b. Types of places,	2	4	2
c. Types of objects or artifacts,	5	3	4
d. Types of events or activities	2	4	4
9. I would use this system again to help me find a photo	2	1	2

The end-user only provided an additional comment only for system B (efe_full): “The system is more accurate than the previous one. Asking for words appearing in the GOAL was enough to get the right picture”

After considering the user feedback the consortium decided to introduce the following criteria to the end-user evaluation framework:

- The assigned time per test set should be reduced (less than five minutes each).
- The names of the systems should be meaningless (say A, B or C). That is, modifying efe_no, efe_full or efe_meaning.
- It will be better to perform a unique questionnaire after finishing the three test sets.
- The order in which the systems are being tested could have an affect with respect the end-user expectations.
- It seems to be enough having three different end-users for the Final Test at EFE. Rotating tasks (News Articles) along the systems we can obtain comparable results for the three different systems.

8 Conclusions

This deliverable reports on the integration of MEANING in the TwentyOne Search engine of Irion, and secondly, the application of the search engine to a real end-user task for two months of captions and pictures from the EFE Fototeca database.

The integration involves the import of wordnets from the MCR for Spanish, English, Catalan, Basque and Italian. It also involves the use of WordNet Domains exported from the MCR and integrated in the word-sense-disambiguation system of Irion Technologies. The integration of the MCR was succesful. The disambiguation resulted in the reduction of 50% of the concepts.

We also described the building of an evaluation framework for the end-user task and the design of the first experiment and the results of a first pilot test, carried out by a single user. This pilot test validated the design of the end-user evaluation. Some small adjustments will be made to the set up of the validation.

A larger automatic test will be carried out on the built indexes with queries in all languages. The results of this test will be reported in deliverable 8.3. The final end-user evaluation will be caried out at the end of the project and will be described in deliverable 8.4.

9 References

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

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

Vossen, P., E. Glaser, H. Van Zutphen, R. Steenwijk, "Validation of MEANING", *WP8.1 Deliverable 8.1*, MEANING, IST-2001-34460, Irion Technologies BV, Delft, The Netherlands. 2004.

Appendix A

Instructions for the user (Pilot Test)

Test 1 of 3

Please, read these instructions carefully.

You are preparing four articles with accompanying pictures. System **A** provides you access to the EFE Fototeca database.

System **A** can be located at http://efe.irion.nl/efe_no

You need Internet Explorer versions 5.0 and higher.

You can login to System **A** using username “tester1” and password “w!mp!ek”.

In particular, you are preparing a news article of TOPIC about CONTEXT. You feel that the text would be well served at this point by a visual showing GOAL. Query the Fototeca database using System **A** for an appropriate photograph.

Revise the results and select the appropriate picture.

When you find an appropriate photograph, click on the button labeled “This is the right picture”. If this picture is not appropriate, click on the button labeled “This is the wrong picture”.

If you do not find an appropriate photo the first time, try modifying your query, adding, removing or changing words from the original query.

You can also select more than one picture for each news article.

However, the total time for locating the appropriate pictures for the four news articles is only **twenty minutes**.

When finishing you will have another **ten minutes** for answering a questionnaire.

Thank you for your help in evaluating this system!

News Article 3

TOPIC = ECONOMIA

CONTEXT = El petróleo está subiendo de precio descontroladamente y hay un debate político internacional sobre las causas y las soluciones.

GOAL = Un político hablando sobre el precio del carburante.

News Article 4

TOPIC = JUSTICIA

CONTEXT = Mejicanos condenados a muerte en Estados Unidos.

GOAL = La familia de un recluso mejicano.

News Article 6

TOPIC = MILITAR

CONTEXT = Se están produciendo unas maniobras militares de la OTAN con un gran despliegue de efectivos en varias regiones.

GOAL = Unos soldados simulando un ataque.

News Article 7

TOPIC = MEDIO AMBIENTE

CONTEXT = Se ha celebrado una reunión sobre el tratamiento judicial de los delitos ecológicos.

GOAL = Un juez especialista en medio ambiente.

Instructions for the user (Pilot Test)

Test 1 of 3

Please, read these instructions carefully.

You are preparing four articles with accompanying pictures. System **B** provides you access to the EFE Fototeca database.

System **B** can be located at http://efe.irion.nl/efe_full

You need Internet Explorer versions 5.0 and higher.

You can login to System **B** using username “tester1” and password “w!mp!ek”.

In particular, you are preparing a news article of TOPIC about CONTEXT. You feel that the text would be well served at this point by a visual showing GOAL. Query the Fototeca database using System **B** for an appropriate photograph.

Revise the results and select the appropriate picture.

When you find an appropriate photograph, click on the button labeled “This is the right picture”. If this picture is not appropriate, click on the button labeled “This is the wrong picture”.

If you do not find an appropriate photo the first time, try modifying your query, adding, removing or changing words from the original query.

You can also select more than one picture for each news article.

However, the total time for locating the appropriate pictures for the four news articles is only **twenty minutes**.

When finishing you will have another **ten minutes** for answering a questionnaire.

Thank you for your help in evaluating this system!

News Article 9

TOPIC = POLITICA

CONTEXT = Se ha nombrado el nuevo gobierno. Hoy ha sido la toma de posesión y se han fotografiado en las escaleras la Moncloa.

GOAL = El gobierno en pleno en las escaleras del Palacio de la Moncloa.

News Article 10

TOPIC = TERRORISMO

CONTEXT = Sigue la violencia en Colombia y especialmente en Medellín. Las muertes son una estampa cada vez mas habitual en las calles de la ciudad.

GOAL = Un entierro en Medellín.

News Article 12

TOPIC = ECONOMIA

CONTEXT = El gobierno ha anunciado un nuevo impuesto municipal.

GOAL = La oficina de un ayuntamiento.

News Article 14

TOPIC = SOCIEDAD

CONTEXT = Se ha conocido el programa de las fiestas de San Isidro en Madrid.

GOAL = Un concejal presentando las fiestas de San Isidro.

Instructions for the user (Pilot Test)

Test 3 of 3

Please, read these instructions carefully.

You are preparing four articles with accompanying pictures. System **C** provides you access to the EFE Fototeca database.

System **C** can be located at http://efe.irion.nl/efe_full

You need Internet Explorer versions 5.0 and higher.

You can login to System **C** using username “tester1” and password “w!mp!ek”.

In particular, you are preparing a news article of TOPIC about CONTEXT. You feel that the text would be well served at this point by a visual showing GOAL. Query the Fototeca database using System **C** for an appropriate photograph.

Revise the results and select the appropriate picture.

When you find an appropriate photograph, click on the button labeled “This is the right picture”. If this picture is not appropriate, click on the button labeled “This is the wrong picture”.

If you do not find an appropriate photo the first time, try modifying your query, adding, removing or changing words from the original query.

You can also select more than one picture for each news article.

However, the total time for locating the appropriate pictures for the four news articles is only **twenty minutes**.

When finishing you will have another **ten minutes** for answering a questionnaire.

Thank you for your help in evaluating this system!

News Article 16

TOPIC = SUCESOS

CONTEXT = Problemas de integración de la comunidad gitana en Andalucía.

GOAL = Un acusado por un altercado entre gitanos.

News Article 17

TOPIC = ECONOMIA

CONTEXT = Un informe subraya el precio cada vez más elevado, la poca oferta y la escasa frescura de los productos agrícolas en los centros comerciales.

GOAL = Unos empleados de un centro comercial preparando el puesto de hortalizas.

News Article 18

TOPIC = DEPORTES

CONTEXT = El ministerio ha decidido aumentar la seguridad en los campos de fútbol. Se harán registros mas rigurosos a los aficionados.

GOAL = Un guardia registrando a unos aficionados en un estadio.

News Article 21

TOPIC = TRANSPORTE

CONTEXT = Se anuncia una huelga de trabajadores de los remolcadores en el Canal de Panamá. Se espera que de llevarse a efecto se genere un caos de trafico marítimo en pocos días.

GOAL = Unas barcas remolcadoras amarradas al muelle del Canal de Panamá

Appendix B

User Evaluation Questionnaire (Pilot Test)

System ___

Please agree or disagree with the following statements (1 = strongly agree, 2 = agree, 3 = have no opinion, 4 = disagree, 5 = strongly disagree):

___ The instructions were clear.

___ I succeeded in getting what I wanted done.

___ The queries were normal and natural.

___ The system understood what I wanted say.

___ At each point during the interaction I understood what I could say.

___ The system behaved as expected.

___ The interaction was very long.

I had particular trouble with queries for:

___ Types of people,

___ Types of places,

___ Types of objects or artifacts,

___ Types of events or activities,

other _____

___ I would use this system again to help me find a photo.

If you have any additional comments, please provide them below. Thank you for your participation in this evaluation.