Project ref. no.	*IST-1999-11438*
Project acronym	**MUCHMORE**
Project full title	Multilingual Concept Hierarchies for Medical Information Organization and Retrieval

Security (distribution level)	*Public*
Contractual date of delivery	*Month 36*
Actual date of delivery	*Month 36*
Deliverable number	*D0.6*
Deliverable title	*Final Report*
Type	*Report*
Status & version	*Final Version*
Number of pages	*37*
WP contributing to the deliverable	*WP0*
WP / Task responsible	*DFKI*
Author(s)	*Paul Buitelaar, Bogdan Sacaleanu, Špela Vintar, Diana Steffen (DFKI); Martin Volk (EIT); Hervé Dejean, Eric Gaussier (XRCE); Dominic Widdows (CSLI); Oktavian Weiser (ZInfo); Robert Frederking (CMU)*
EC Project Officer	*Yves Paternoster*
Keywords	*Cross-Lingual Information Retrieval; Medical Domain; Semantic Annotation; Sense Disambiguation; Relation Extraction; Classification-based Retrieval; Summarization; EBT; PRF; MeSH; UMLS; EuroWordNet*
Abstract (for dissemination)	*Within the scope of establishing a Cross-Lingual Information Retrieval framework, the MuchMore project pursued the following aims: Research regarding the effective combination of different approaches and heterogeneous resources and their integrated use for multilingual information access and management, including performance evaluation for realistic information access tasks; Research and technology development concerning the automated acquisition of domain-specific linguistic resources and effective use of multilingual concept hierarchies; Demonstration of a cross-lingual information access prototype system for the medical domain, that provides access to multilingual information on the basis of combined use of corpus analysis and a domain-ontology.*

1 Introduction

MuchMore provides a framework for integrating and refining existing technologies and developing new approaches to cross-lingual information retrieval for the medical domain. Existence of very large ontologies of domain concepts and extensive corpora for the medical domain has grounded the work toward refinement, integration and comparison of concept-based retrieval methods and corpora-based approaches.

Within the scope of establishing a cross-lingual information retrieval framework, the MuchMore project pursued the following aims:

- Research regarding the effective combination of different approaches and heterogeneous resources and their integrated use for multilingual information access and management, including performance evaluation for realistic information access tasks.

- Research and technology development concerning the automated acquisition of domain-specific linguistic resources and effective use of multilingual concept hierarchies.

- Demonstration of a cross-lingual information access prototype system for the medical domain, that provides access to multilingual information on the basis of combined use of corpus analysis and a domain-ontology.

Along these lines, innovative developments have been done in the following areas:

Combination of Heterogeneous Resources and Methods

Methods and resources already available, as well as within the project new developed, have been integrated for multilingual information management and access. Performance testing provided quantitative feedback on the performance of different methods and resources. Methods used include concept-based approaches (semantic annotation of terms and relations, including disambiguation and filtering), corpus-based approaches (similarity thesauri, example-based translation, pseudo-relevance feedback and classification) as well as combinations of these.

Use of Concept Hierarchies in Cross-Lingual Information Retrieval

Use of medical conceptual structures for mapping both user queries and documents into an intermediate representation, bridging the cross-language barrier, has been investigated and new techniques for automatic semantic annotation have been developed. Concept hierarchies have been integrated in the query processing functionality, which allows for interactive refinement of user queries.

Multi-Document Summarization

Multi-document summarization has been developed and integrated into the cross-lingual retrieval engine. The approach is based on the Maximal Marginal Relevance method of extracting the most relevant and diverse passages from a text (Carbonell and Goldstein, 1998).

2 The MuchMore Prototype

The MuchMore prototype[1] is a cross-lingual document retrieval system that enables users to retrieve documents (in English and/or German), which are relevant to a given query document (in English or German). In the current version of the MuchMore system, query documents are assumed to be German electronic patient records and documents to be retrieved are medical scientific abstracts in both German and English. The MuchMore prototype is implemented as a meta-search engine that provides access to a merged/ranked list of relevant documents from three different search engines and a query construction tool that provides a user interface for extracting and refining structured queries.

2.1 Cross-Lingual Meta-Search Engine (CMU, CSLI, DFKI, EIT)

Within the project, the cross-lingual information retrieval task has been approached from a number of different views, corresponding to (combinations of) concept-based and corpora-based methods: CMU (EBT: Example-Based Translation and PRF: Pseudo Relevance Feedback methods), CSLI (concept-space model), DFKI/EIT (semantic annotation).

Along these lines, three demo systems were developed that have been integrated into a meta-search engine with a common user interface and results presentation:

- A cross-lingual document retrieval system[2] based on semantic annotation, which is using **concept-based** methods for generating an intermediary representation of both queries and documents.
- A cross-lingual thesaurus generation and document retrieval system[3], which uses a **corpora-based** statistical analysis of the distribution of terms in translated documents to generate a representation of terms and documents from both languages in the same mathematical space.
- A cross-lingual document retrieval system[4] leveraging parallel training **corpora** in the relevant languages, employing two methods: one a pseudo-relevance feedback (PRF) approach, and the other an example-based thesaurus (EBT) approach, including a thesaurus generator.

[1] http://lit.dfki.uni-sb.de:8000/prototype/index.html
[2] http://www4.eurospider.ch/cgi-bin/MUCHMORE/QueryCGI?ds=MUCHMORE_pull&sr=0&ri=en&lang=en&f=1&fr=1&n=10
[3] http://infomap.stanford.edu/bilingual
[4] http://nyc.lti.cs.cmu.edu:8080/

2.2 Query Construction Tool (DFKI)

The entry point to the MuchMore prototype is a query construction tool that provides a user interface for extracting and refining structured queries (see Figure 1. below). For this purpose, the following information is displayed:

- the text of the query, serving as reference context for any further refinements
- a list of automatically extracted medical concepts along with their frequency and the semantic relations holding among the concepts
- a browsing option that helps the user to navigate through the concept space and include more general or more specific concepts in the constructed query

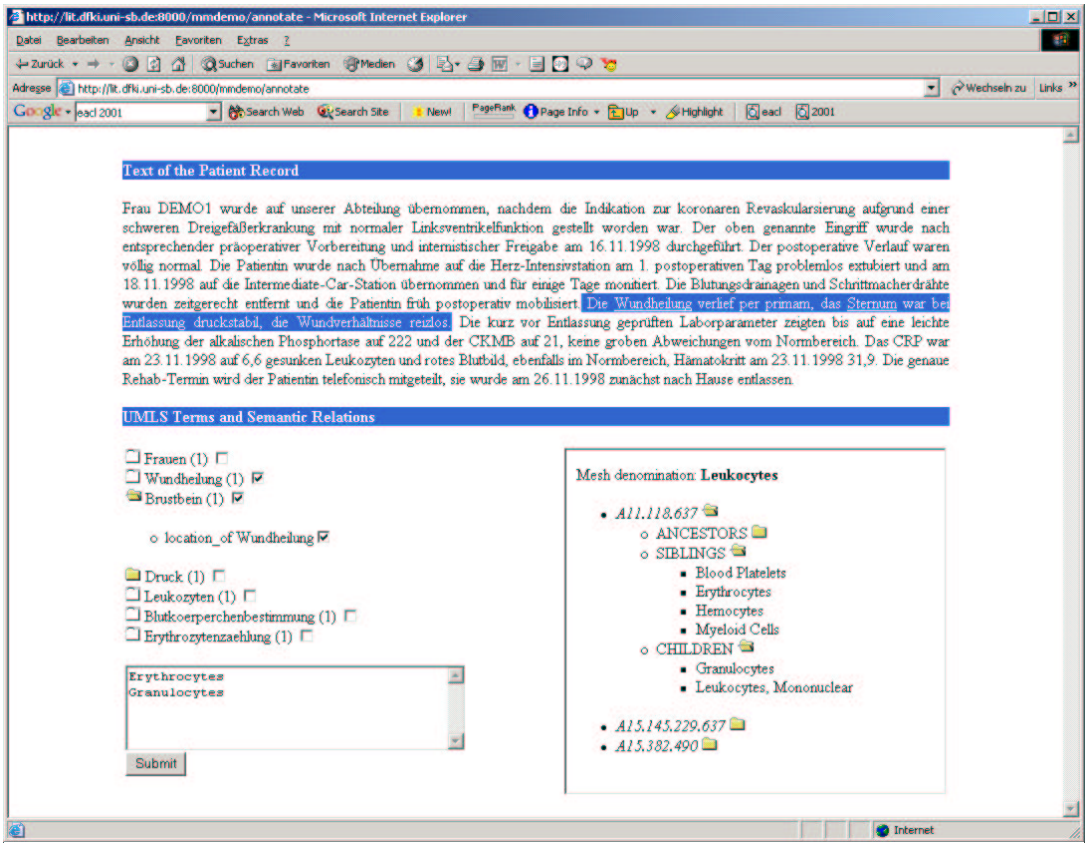

Figure 1: Query Construction Interface

The concept list consists of preferred names of the matched terminology, as found in the controlled vocabulary. Furthermore, on clicking the frequency number associated with a concept, all its instances in the query are highlighted. Thereby the user is not only presented with a normalized medical terminology, according to the controlled vocabulary, but he can also inspect which terms in the query document are instances of which concepts. A list of semantic relations that hold between co-occurring concepts is displayed for each concept. When the user clicks on a listed relation, the context of the relation and its concepts are highlighted, helping the user to make an informed choice on the relevance of the automatically extracted relation.

To allow further query expansion, the prototype provides a browse-able view of the concept hierarchy. By selecting any concept in the extracted list of concepts, an overview is given of

ancestor, sibling and child concepts. By double-clicking any of these, the query can be extended in a way that is relevant to the user needs, with the added concepts shown in a text area below the original concept list.

Once the query has been refined according to the user needs, the underlying information about tokens, lemmas, concept codes and their relations is sent to the selected retrieval engine(s).

2.3 Summarization Tool (CMU)

CMU's web-based demonstrator provides a threefold convenience of summarization when returning results to the user. Flagship summarization is provided through a Maximal Marginal Relevance (MMR) based facility that extracts the most relevant and diverse passages across all returned documents in each language (Carbonell and Goldstein, 1998). This allows the user, at a glance, to judge the both the breadth and tenor of the search engine's results. Further, although the summary is initially constructed across all results, the demonstrator permits the user to specify a subset of documents in order to obtain a summary over one or more particularly interesting documents. For the purposes of being able to provide summaries drawn from multiple documents, the basic MMR implementation has been adapted to work with individual texts or with clusters of topical-related texts. As noted, such summaries are provided for each language present in the results; this adaptation does not extend to cross-lingual cases.

Additionally, each document returned by the search engine and listed for the user is accompanied not just by the document's title, but also an excerpt from the body. This allows the user to get a sense of the contents of each returned document just by perusing the list, without having to explicitly request each. Finally, a summarization by way of grouping like documents by topic is inherent when retrieving results in categorically hierarchical format. By noting where in the subject tree documents reside, the user is able to determine -- without need for reading and assessing document texts -- their topical content.

The input / output behavior for all of these methods is straightforward. In each case, the input is a document ID (or a set of document IDs, in the case of MMR summarization) and the output is the summary text or topic label.

3 Approaches to Cross-Lingual Information Retrieval

3.1 Concept-Based Approach (CSLI, DFKI, EIT, XRCE)

One of the primary goals of the MuchMore project is to develop and evaluate methods for the effective use of multilingual thesauri in the semantic annotation of English and German medical texts and subsequently to evaluate and compare the impact of such semantic information for the purpose of cross-lingual information retrieval (CLIR). In particular, work on semantic annotation with (domain-specific: UMLS[5] and general language: EuroWordNet[6]) semantic resources, based on linguistic analysis of domain-specific corpora (i.e. the underlying document collection) is central to the concept-based CLIR approach described in this section.

[5] http://www.nlm.nih.gov/research/umls/
[6] http://www.illc.uva.nl/EuroWordNet/

3.1.1 Semantic Resources Used

The essential part of any concept-based CLIR system is the identification of terms and their mapping to a language-independent conceptual level.

UMLS

Our basic resource for semantic annotation is UMLS, which is organized in three parts.

The **Specialist Lexicon** provides lexical information for medical terms: a listing of word forms and their lemmas, part-of-speech and morphological information.

Second, the **Metathesaurus** is the core vocabulary component, which unites several medical thesauri and classifications into a complex database of concepts covering terms from 9 languages. Each term is assigned a unique string identifier, which is then mapped to a unique concept identifier (CUI). For example, the entry for *HIV pneumonia* in the Metathesaurus main term bank (MRCON) contains (among others) the concept identifier, the language of the term and the string:

```
C0744975 | ENG | HIV pneumonia
```

In addition to the mapping of terms to concepts, the Metathesaurus organizes concepts into a hierarchy by specifying relations between concepts. These are generic relations like *broader than, narrower than, parent, sibling* etc. Another component of the Metathesaurus provides information about the sources and contexts of the concepts. The UMLS 2001 version includes 1.7 million terms mapped to 797,359 concepts, of which 1.4 million entries are English and only 66,381 German. Only the MeSH[7] (Medical Subject Heading) part of the Metathesaurus covers both German and English, therefore we only use MeSH terms for corpus annotation.

The third part is the **Semantic Network**, which provides a grouping of concepts according to their meaning into 134 semantic types. The concept above would be assigned to the class *T047, Disease or Syndrome*. The Semantic Network then specifies potential relations between those semantic types. There are 54 hierarchically organized domain-specific relations, such as *aspects, causes, location of* etc.

We assign semantic codes to each sentence based on the linguistic information. MeSH codes were assigned to documents and to queries. UMLS concept identifiers were used as the basis for finding semantic relations.

EuroWordNet

EuroWordNet is a multilingual database with WordNets for a large number of European languages (Vossen, 1997). In addition to annotation with UMLS, terms are annotated also with EuroWordNet to compare domain-specific and general language use. EuroWordNet is a multilingual database for several European languages and is structured in similar ways to the Princeton WordNet (Fellbaum, 1997). Each language specific (Euro)WordNet is linked to all of the others through the so-called Inter-Lingual-Index (ILI), which is based on WordNet1.5. Via this index the languages are interconnected, so that it is possible to move from a word in one language to similar words in any of the other languages in the EuroWordNet database. For our current purposes we use only the German and English parts of EuroWordNet.

[7] http://www.nlm.nih.gov/mesh/meshhome.html

All information in (Euro)WordNet is centered around so-called synsets, which are sets of (near-) synonyms. The different senses of a term are therefore simply all the synsets that contain it. The goal of disambiguation is to narrow down these possibilities, ideally to a single sense. A term can be simple (*man*) or complex (*rock_and_roll*). A synset is identified by a unique identifier, called offset. Because meanings between languages cannot be exactly mapped one-to-one, there may be more than one synset within a language that is mapped on the same concept in the ILI. In order to distinguish between these, every synset was given a unique identifier (ID)[8], as shown in Table 1-1:

	Offset - ID	Synset
German	3824895 - 1	Fingergelenk
	3824895 - 2	Fingerknochen
	3824895 - 3	Knöchel
English	3824895	knuckle, knuckle joint, metacarpophalangeal joint

Table 1: EWN Example

3.1.2 Linguistic and Semantic Annotation (DFKI)

Identification and annotation of terms with concepts and relations is based on linguistic analysis: part-of-speech, morphology and phrases (chunks). Results of linguistic and semantic (terms, semantic relations) annotation are integrated in a multi-layered XML format, which organizes various levels as separate tracks with options of reference between them via indices[9].

he aim was to design an annotation format that would include all layers and adequately represent relationships between them, while at the same time remaining logical and readable, efficient for parsing and indexing as well as flexible for future additions and adjustments (Vintar et al. 2002).

We will explain the annotation format with the following example sentence from an abstract in the field of psychiatry.

Balint syndrom is a combination of symptoms including simultanagnosia, a disorder of spatial and object-based attention, disturbed spatial perception and representation, and optic ataxia resulting from bilateral parieto-occipital lesions.

Each document is split into sentences and the XML annotation is based on them. Each `<sentence>` contains a `<text>` block that holds the tokens as XML content, and both lemma and part-of-speech information as XML attributes.

```
<text>
   <token id="w1" pos="NN"> Balint </token>
   <token id="w2" pos="NN"> syndrom </token>
   <token id="w3" pos="VBZ" lemma="be"> is </token>
   <token id="w4" pos="DT" lemma="a">  a </token>
```

[8] In our case only for German, as the English synsets correspond to the ILI directly.

[9] For further details on the annotation format and process, please refer to deliverable D4.1. A demo version of the automatic (linguistic/semantic) annotation system is available at: http://muchmore.df.de/demo2.html In addition, an interactive GUI is available (MMV: "MuchMore Viewer") for annotation development purposes (i.e. to check the validity of the automatic annotation and to make corrections interactively).

```
   <token id="w5" pos="NN" lemma="combination"> combination </token>
   <token id="w6" pos="IN" lemma="of"> of </token>
   <token id="w7" pos="NNS" lemma="symptom"> symptoms </token>
   ...
   <token id="w20" pos="JJ" lemma="spatial"> spatial </token>
   <token id="w21" pos="NN" lemma="perception"> perception </token>
   <token id="w22" pos="CC" lemma="and"> and </token>
   <token id="w23" pos="NN" lemma="representation"> representation </token>
   ...
</text>
```

Linguistic analysis determines noun phrases, adjective phrases and prepositional phrases. In this example it determines - among others - a noun phrase (NP) for words `w1` and `w2` *Balint syndrom* and a more complex noun phrase from `w20` to `w23` *spatial perception and representation*.

```
<chunk id="c1" from="w1" to="w2" type="NP"/>
<chunk id="c7" from="w20" to="w23" type="NP"/>
```

In addition, each `<sentence>` contains semantic annotations. In a first block we store pointers to EuroWordNet (EWN) senses. For the example sentence we determined that word `w21`, *perception*, has four EWN senses, related to *perceiving - sensing*, *perception*, and *perceptual experience*.

```
<ewnterm id="e5" from="w21" to="w21">
   <sense offset="487490"/>
   <sense offset="3890199"/>
   <sense offset="3955418"/>
   <sense offset="4002483"/>
</ewnterm>
```

At the core of semantic annotation are UMLS terms and MeSH codes. For the example sentence the words `w20` and `w21` point to the concept with a preferred name ”Space Perception”, which corresponds to the CUI code C0037744 and TUI code T041 (i.e. “Mental Process”). In addition, this concept is linked to two MeSH codes, which stand for two positions of the term ”Space Perception” in the MeSH tree of concepts, the first under the node ”Perception” and the second under ”Visual Perception”. Finally, word `w26` (*optic*) triggered the concept ”Optics” (with one corresponding MeSH code).

```
<umlsterm id="t7" from="w20" to="w21">
   <concept   id="t7.1"   cui="C0037744"   preferred="Space   Perception"
tui="T041">
     <msh code="F2.463.593.778"/>
     <msh code="F2.463.593.932.869"/>
   </concept>
</umlsterm>

<umlsterm id="t8" from="w26" to="w26">
   <concept id="t8.1" cui="C0029144" preferred="Optics" tui="T090">
     <msh code="H1.671.606"/>
   </concept>
</umlsterm>
```

The most specific information is on the semantic relations that are derived from the UMLS Semantic Network. For example, it indicates that ”Space Perception” is an issue in ”Optics” which is coded in the following manner. Note that the XML attributes `term1` and `term2`

point to the UMLS concepts introduced in the example above.

```
<semrel id="r7" term1="t7.1" term2="t8.1" reltype="issue_in"/>
```

3.1.3 Sense Disambiguation (CSLI, DFKI)

Obviously, terms may correspond to more than one concept in the semantic resources used, which is of particular importance in the CLIR context. For instance, the English word *drug* when referring to medically therapeutic drugs would be translated as *medikamente*, while it would be rendered as *drogen* when referring to a recreationally taken narcotic substance of the kind that many governments prohibit by law.

The ability to disambiguate may therefore be crucial to applications such as CLIR, since search terms entered in the language used for querying must be appropriately rendered in the language used for retrieval. Because of this potential importance to cross-lingual language and information applications, sense disambiguation has been one of the areas of focus of the MuchMore project.

Evaluation Corpora

An important aspect of sense disambiguation is the evaluation of different methods and parameters. Unfortunately, there is a lack of test sets for evaluation, specifically for languages other than English and even more so for specific domains like medicine. Given that the focus of the project is on German as well as English text in the medical domain, we had to develop a number of manually annotated evaluation corpora (lexical samples[10]) to test the different disambiguation methods developed within the project with EuroWordNet (or rather GermaNet) for German, and with UMLS for both German and English.

To support manual annotation we developed a lexical sample annotation tool based on the ANNOTATE tool that has been developed in the context of the NEGRA project on syntactic annotation (Plaehn and Brants, 2000). In selecting an ambiguous occurrence to be manually annotated (i.e. disambiguated), the annotator is presented with the extended context (left/right neighbor sentences) and the senses for this particular word. By selecting one or more of these, the annotator tags every occurrence of the word with the appropriate sense(s). If the lexical semantic resource does not contain an appropriate sense for the corresponding context, the annotator can choose to annotate with *unspec* (unspecified). To further assist the annotator, there is access also to corresponding hierarchies (hypernymy in GermaNet or broader term in UMLS).

Selection of ambiguous terms for the GermaNet evaluation corpus proceeds by compiling a list of terms with high domain relevance, at least 100 occurrences in the medical corpus and with more than one sense in GermaNet. From this list we selected 40 terms, for each of which we then automatically extracted 100 occurrences at random. Three annotators, a medical expert and two linguistics students, were assigned the task of annotating the selected occurrences for these ambiguous terms. We also employed non-experts, as they would not have much difficulty in tagging occurrences in a medical corpus, because most of the terms express rather commonly known (medical or general) concepts.

The process of selecting terms for the UMLS evaluation corpora (English and German) is based on automatically generated lists of ambiguous UMLS terms. From these we selected a

[10] See (Kilgarriff, 1998) for a discussion of lexical sample corpora for the evaluation of sense disambiguation.

set of 70 frequent terms for English (token frequencies at least 28, 41 terms having token frequencies over 100). For German, only 24 terms could be selected (token frequencies at least 11, 7 terms having token frequency over 100[11]), as the German part of UMLS (or rather MeSH) is rather small. The level of ambiguity for these UMLS terms is mostly limited to only 2 senses; only 7 English terms have 3 senses. In the case of UMLS, medical experts were involved in the manual annotation, two for the German part and three[12] for the English part.

In order for an automatic system to decide which sense is more appropriate in a given context, it is a prerequisite that at least human annotators agree between them on this. We therefore computed the inter-annotator agreement (IAA) between the annotators involved in the various manual annotation tasks[13]. The IAA for the GermaNet evaluation corpus varies from very low to very high, but is on average at 70%. The agreement scores for the UMLS evaluation corpora vary also highly, with an average of 65% for German and 51% for English.

Methods

Methods for disambiguation can effectively be divided into those that require manually annotated training data (supervised methods) and those that do not (unsupervised methods). In general, supervised methods are less scalable than unsupervised methods because they rely on training data, which may be costly and unrealistic to produce, and even then might be available for only a few ambiguous terms. The goal of our work on disambiguation in the MuchMore project is to enable the correct semantic annotation of entire document collections with all terms, which are potentially relevant for organisation, retrieval and summarisation of information. Therefore a decision was taken early on in the project to focus on unsupervised methods, which have the potential to be scaled up enough to meet our needs. The methods developed are: bilingual, dictionary-based, domain-specific and instance-based learning[14].

Bilingual methods take advantage of having a translated corpus, because knowing the translation of an ambiguous word can be enough to determine its sense. Dictionary based methods use relations between terms as deduced from a dictionary or some other semantic resource to determine which sense is being used in a particular instance. Domain-specific methods use the fact that certain meanings of general terms are far more important than others in specific domains (for example, in the medical domain, “operation” is far more likely to refer to a surgical operation than a military operation). Instance-based learning is a machine-learning technique for classification that an also be applied in sense disambiguation if each sense is treated as a class and an ambiguous occurrence as an instance to be classified.

Results

The best results for precision ranged from 74% (English) to 79% (German), achieved by the dictionary-based method on the UMLS evaluation corpora with a coverage of 83% (English) and 87% (German). All other methods scored mostly lower than this, although the domain-specific method achieved a precision of 77%-99% on the GermaNet evaluation corpus but with very low coverage.

[11] We automatically created evaluation corpora using a *random selection* of occurrences if the term frequency was higher than 100, and using *all* occurrences if the term frequency was lower than 100.
[12] We had two German annotators and an American annotator. The German ones annotated both the German and the English UMLS evaluation corpora, while the American annotator participated in only the English UMLS evaluation corpus.
[13] The importance of inter-annotator agreement has been discussed in detail in (Kilgarriff 98).
[14] For more details on each of the methods, please refer to deliverable D5.1

While none of the methods required manually annotated training data, the more precise methods relied on other sources of knowledge for their success. In particular, the dictionary-based method made use of the detailed structure of UMLS. The bilingual method relied on the availability of a parallel corpus – it would be impractical to construct these resources purely for the sake of disambiguation. On the other hand, the domain-specific and instance-based learning methods were less resource intensive, using only the sense inventory of GermaNet and domain specific corpora for training.

Evaluation of Sense Disambiguation in CLIR

The dictionary-based method, which performed well on the UMLS evaluation corpora, was used to automatically disambiguate the whole Springer corpus. Experiments were then carried out to assess the contribution of automatic semantic annotation and disambiguation to document retrieval using the CSLI search engine, compared with a baseline model which indexed documents only by their tokens.

The results (Table 2 and Table 3) showed that semantic annotation, with and without automatic disambiguation, gave better results than using only tokens (especially for German, where the tokens only model is particularly naïve). Disambiguation had little effect on the precision for the top few retrieved documents, but enabled the search engine to retrieve more relevant documents further down the list of results, as reflected by the consistently higher scores for mean average precision obtained on the disambiguated corpora.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
Tokens only	0.100	228	0.258	0.252
Automatic Annotation	0.112	242	0.361	0.276
Automatic Annotation and Disambiguation	0.116	254	0.339	0.288

Table 2 English results with disambiguation

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
Strings only	0.079	210	0.151	0.148
Automatic Annotation	0.110	241	0.299	0.292
Automatic Annotation and Disambiguation	0.113	243	0.290	0.292

Table 3: German results with disambiguation

Cross-lingual retrieval runs (English to German and German to English) produced very similar conclusions, confirming the hypothesis that disambiguation helps the search engine to maintain consistently higher precision as we go further down the list of retrieved documents.

3.1.4 Term Extraction (XRCE)

When working on specialized languages and specific domains, terminology plays a crucial role, inasmuch as it aims at describing and organizing the knowledge of the domain through the concepts, and their lexical realizations, that are used. In using multilingual thesauri (i.e. UMLS, MeSH) for CLIR purposes such as within the MuchMore project, the additional step of linking terms across languages is required. In specialized domains, parallel texts, i.e. texts in two or more languages, which are translation of each other, provide an ideal material to follow, from a multilingual point of view, the evolution of the terminology of a domain, and

to update existing resources. Therefore, the work on term extraction had a focus on the task of enriching a domain specific thesaurus through bilingual lexicon extraction from parallel corpora. The thesaurus used is the Medical Subject Headings (MeSH), and its German version DMD, available through the Unified Medical Language System (UMLS). The main goal of the research has been to show that one can develop methods for bilingual lexicon extraction, from parallel and comparable corpora, and tools, based on the extracted lexicons, first to help terminologists enrich existing thesauri, and second to help users find relevant information in different languages.

Description of the Data

We used several kinds of resources for the term extraction experiments: a corpus, a general bilingual lexicon and a specialized thesaurus. The corpus is based on the MuchMore Springer collection of medical abstracts extracted from the Springer LINK website[15]. These abstracts are "partial" translations of each other, because in some cases the English writer directly summarizes the articles in English, rather than translating the German abstracts.

The set of abstracts is used both as a comparable corpus, in which case we do not make use of alignment information, and as a parallel corpus. There is a continuum from parallel corpora to fully unrelated texts, going through comparable corpora. The comparable corpus we use is in a way "ideal" and is biased in the sense that we know the translation of a German word of the German corpus to be, almost certainly, present in the English corpus. However, this bias, already present in previous works, does not impact the comparison of the methods we are interested in, all methods being equally affected. Indeed, the results we obtain with the standard method are in the range of those reported in previous works.

As a general bilingual resource, we use the German/English ELRA dictionary[16], which contains about 50,000 bilingual entries. The medical thesaurus used is MeSH (Medical Subject Headings), and its German version, DMD, provided by DIMDI[17]. Both thesauri were extracted from UMLS, through which the MeSH English entries and the DMD German entries are aligned, so we could extract a bilingual thesaurus. Since DMD is smaller than MeSH, the resulting bilingual thesaurus only contains 15,000 bilingual entries, when MeSH contains 200,000 entries.

Term Extraction from Parallel Corpora

A parallel corpus is a bilingual corpus the elements of which are translations of each other, and which is aligned, usually through an automatic procedure, at the sentence level (which means that each sentence in one language is associated with its translation in the other language). The methodology we follow to extract bilingual lexicons made of single and multi-word units corresponds to a parse-parse strategy, and is based on three steps[18]:

1. Word alignment across languages,
2. Term extraction in both languages,
3. Term alignment across languages, based on outputs of steps 1 and 2.

[15] http://muchmore.dfki.de/resources1.htm
[16] http://icp.grenet.de/ELRA/home.html
[17] http://dimdi.de
[18] For a detailed account of the method used, please refer to the XRCE report on term extraction (D7.1)

Term Extraction from Comparable Corpora

Bilingual lexicon extraction from non-parallel but comparable corpora relies on the assumption that if two words are mutual translations, then their more frequent collocates (taken here in a very broad sense) are likely to be mutual translations as well. Based on this assumption, a standard approach consists in building context vectors, for each source and target word, which aim at capturing the most significant collocates. The target context vectors are then translated using a general bilingual dictionary, and compared with the source context vectors.

The use of a general bilingual dictionary to translate context vectors is justified by the fact that if the context vectors are sufficiently large, then some of their elements are likely to belong to the general language and to the bilingual dictionary, and we can thus expect the translated context vector of word *t* to be, in average, closer to the context vector of the translation *s* of *t*. It has to be noted that the above strategy makes sense even when *t* is present in the bilingual dictionary, since the corpus may display a particular, technical usage of *t*.

Results

Evaluation of bilingual term extraction from parallel corpora gives a F1-score of 85% when the 10 best candidates are taken into account. A gold standard bilingual lexicon was extracted from the Springer corpus for this evaluation (1800 words).

#best candidates	Precision	Recall
1	56.52	50.98
2	71.01	64.05
5	84.78	76.47
10	89.85	81.04

Table 4: Evaluation of term alignment with the parallel corpus

In bilingual lexicon extraction from comparable corpora, we aimed at combining information provided by three different probabilistic lexical translation models, the first one mainly based on the corpus, the second one mainly based on a multilingual thesaurus, and the third one derived from a bilingual dictionary. Special attention has been given to the use of multilingual thesauri, and different search strategies based on such thesauri have been investigated. A method to optimally combine the different resources for bilingual lexicon extraction was developed and extended to cross-language information retrieval where the bilingual lexicon is not a goal per se but rather a way to retrieve documents in different languages. Our results show that the combination of the resources significantly improves results, and that the use of the hierarchical information contained in our thesaurus, UMLS/MeSH, is of primary importance.

Methods	p=5	p=10
Viterbi	71.3/14.7	79.7/14.7
Complete(100)	75.4/14.1	80.3/14.1
Complete(200)	75.4/12.3	83.2/12.3
SubTree 10	75.8/11	82.4/11
SubTree 20	76.4/11.7	84.1/11.7
SubTree 50	77.3/11.2	83.6/11.2
SubTree 100	76.9/11.8	83/11.8

Table 5: Evaluation of different search strategies (p: # of best candidates)

Thesaurus enrichment

The first extension we address is the introduction of new strings (following UMLS terminology) associated with a concept in the thesaurus. If one element of the bilingual extracted lexicon is in the thesaurus, the translated candidate can be directly proposed as a possible addition to the part of the thesaurus corresponding to its language. Such new strings usually correspond to synonyms as well as spelling or term variants. For example, the German string *Karzinom* is associated with the UMLS concept C0007097. The corresponding English string in UMLS is *carcinoma*. Through the term alignment process, we can propose a new German string: *Carcinom*. Note that this spelling difference is in fact due to two different German spelling used in medical texts. New strings corresponding to morpho-syntactic variations can also be detected. Thus, our alignment provides a new string for the entry *Lebertransplantation*: *Transplantion der Leber*, which was not part of the original entry which contains: *Lebertransplantation*, *Hepar-Transplantation*, *Transplantation*, *Hepar-*, *Leber-*. A second kind of enrichment is the addition of new concepts in one language. In some cases no German string is proposed for a given concept class. For example, the German thesaurus has no associated strings with C0334281 (*malignant insulinoma*) and C0406864 (*flap loss*). Our alignment tool allows us to propose the following candidates: *malignen Insulinom* for C0334281, and *Lappenverlust* for C0406864, propositions that a terminologist can review before deciding to enter them or not in the thesaurus. From the 700 medical abstracts that our corpus contains, about 1400 new German terms can be proposed in such a way to the terminologist.

The most difficult situation is the addition of new concepts and associated strings in the two languages. In this case, for a given pair of aligned terms, we can propose to the terminologist in charge of updating the thesaurus the set of concept classes, which are closest to the pair of terms under consideration. The terminologist will then decide whether or not to create new concept classes in the thesaurus. This strategy is valid when the concept described by the pair of terms is close to existing concepts, for example when it is a narrower concept of an existing concept. For instance, no concept exists in UMLS for the German string *chronische Pankreatitis* or its English translation, *chronic pancreatitis*. Computing a similarity between concepts and these terms yields C0030305 (*pancreatitis*) as the closest concept class to both candidate terms. In general, if the term is composed with some words present in the thesaurus, the list of concepts proposed to the terminologist is relevant.

Nevertheless, if none of the words appearing in the new terms is present in the thesaurus, the above strategy will fail. A solution in this case is to combine hierarchical information with some particular morpho-syntactic patterns. For example, words with the suffix *-ectomy* tend to occur below the concept C0543467 (*Surgical Procedures*) and words with the suffix *-graphy* tend to occur below *Diagnosis*. Through this information, we can propose new concepts to be added to the thesaurus, as well as target the subpart of the thesaurus closest to this new concept.

3.1.5 Relation Extraction (DFKI)

Semantic relations are annotated on the basis of the UMLS Semantic Network, which defines binary relations between semantic types (TUIs) in the form of triplets, for example `T195 - T151 - T042` meaning *Antibiotic - affects - Organ or Tissue Function*. We search for all pairs of semantic types that co-occur within a sentence, which means that we can only annotate relations between items that were previously identified as UMLS terms.

According to the Semantic Network relations can be ambiguous, meaning that two concepts may be related in several ways. For example:

Diagnostic Procedure	*assesses_effect_of*	*Antibiotic*
Diagnostic Procedure	*analyzes*	*Antibiotic*
Diagnostic Procedure	*measures*	*Antibiotic*
Diagnostic Procedure	*uses*	*Antibiotic*

Since the semantic types are rather general (e.g. *Pharmacological Substance, Patient or Group*), the relations are often found to be vague or even incorrect when they are mapped to a document. If for example the Semantic Network defines the relation *Therapeutic Procedure -- method_of -- Occupation or Discipline*, this may not hold true for all combinations of members of those two semantic classes, as seen in **discectomy -- method_of -- history*. Given the ambiguity of relations and their generic nature, the number of potential relations found in a sentence can be high, which makes their usefulness questionable. A manual evaluation of automatic relation tagging in a small sample by medical experts showed that only about 17% of relations were correct, of which only 38% were perceived as significant in the context of information retrieval.

On the other hand, many relations present in our texts are not identified by automatic relation tagging. One possible reason for this may be the incompleteness of the Semantic Network, but a more accurate explanation is that relationships are constantly being woven between concepts occurring together in a specific context, thus creating novel or unexpected links that would not exist between concepts in isolation.

For the above reasons we developed methods to deal with each of the problems described, relation filtering and relation extraction.

Relation Filtering

The first task was to tackle relation ambiguity, i.e. to select correct and significant relations from the ones proposed by automatic UMLS lookup. The method is composed of two steps following two initial hypotheses:

- Interesting relations will occur between interesting concepts.
- Relations are expressed by typical lexical markers, such as verbs.

Following our first hypothesis we expect interesting and true relations to occur between items that are specific rather than general, and thus not too frequent. To measure this specificity we use the inverse document frequency (IDF) of the concept's code (CUI), which assigns a higher weight to concepts occurring only in a subset of documents in the collection. We decided to use IDF instead of the generally used TF-IDF, because term frequency (TF), if multiplied with IDF, will assign a higher score to frequent terms like *patient, therapy, disease*. Relations between items with the IDF weight below a certain value are removed.

Relations may be represented by various linguistic means (i.e. lexical markers). In a rule-based approach such markers would be specified manually, however we chose to use a co-occurrence matrix of verbs and automatically tagged relations. This is based on the assumption that some verbs are more likely to signify a certain relation than others. The co-occurrences are normalized and non-lexical verbs filtered out, so that for each lexical verb we

get a list of relations it most likely occurs with. This information is then used to remove relations that occur with an untypical verb.

Extraction of New Relation Instances

The identification of new instances of relations was based on observed co-occurrences of concepts, where instead of the semantic types (TUI) from the Metathesaurus we use MeSH classes. This gives us flexibility in choosing the number of semantic classes, depending on the level in the hierarchy. We use co-occurrences on the second level, meaning that we strip full MeSH codes assigned to each concept to only the top node letter and first order children. For each UMLS semantic relation we then compute a list of typical MeSH pairs, for example:

treats - D27|C23, D3|C23, E7|C23, E7|C2, ...

Once these patterns of correspondence between pairs and relations are established, we may extract new instances of relations on the basis of co-occurring MeSH codes within the sentence. The extraction method can be tuned in terms of precision and recall by setting the MeSH-pair frequency threshold. For our current document collection and CLIR purposes this was set to 150.

Evaluation

The main goal of the experiments that we describe was to evaluate the usefulness of semantic relations in CLIR, where we explore the possibilities of modifying and expanding existing semantic resources, i.e. UMLS. Unfortunately, due to low term coverage for German, only very few semantic relations were found on the query side, and it was therefore impossible to assess their value. For this reason we opted for a monolingual approach, using English queries over the English document collection, however without indexing tokens and lemmas but relying solely on semantic information. In the tables below we present the retrieval results in four columns: mean average precision (mAP), absolute number of relevant documents retrieved (RD), average precision at 0.1 recall (AP01) and precision for the top 10 documents retrieved (P10).

We compare five runs (results in the table below). The first has UMLS-based relations filtered with the IDF method (umls_idf_filt), the second was additionally filtered with the verb method (umls_idf_vb_filt). We then introduce newly extracted relation instances, first to the filtered version of the corpus (umls_idf_vb_new), then to the baseline UMLS-annotated version (umls_new) and finally, we annotate relations using only our method for extracting new relation instances (only_new).

	mAP	RD	AP01	P10
umls_idf_filt	0.126	203	0.315	0.280
umls_idf_vb_filt	0.107	175	0.282	0.264
umls_idf_vb_new	0.124	197	0.336	0.308
umls_new	0.153	259	0.419	0.344
only_new	0.116	213	0.363	0.280

Table 6: Results of relation filtering and extraction indexing relations only

The results show that each filtering step significantly decreases both recall and precision, while adding new relations -- as we would expect -- works well. The highest precision and recall were achieved with a combination of UMLS annotation and new relations, and this combination also outperforms the baseline.

3.2 Hierarchical MeSH Concept Classification (CMU)

3.2.1 Overview

The MeSH hierarchy is used in two places in MuchMore: as a translation/retrieval aid in the search engine, and as a representation aid in the web demonstrator. Both cases rely on the documents in the search space being labeled in accordance with the MeSH hierarchy. For this task, the OHSUMED-87 corpus was used as training data for a machine-based automatic assignment. Since the OHSUMED-87 corpus is in English, the process for labeling the English half of the search space is straightforward. In order to label the German half of the search space, the aforementioned parallel training corpora were employed as a conduit; first, the English training corpus is categorized by the system, then the labelings are transferred to the German part of the training documents, and, finally, the labeled German training corpus is used to label the German half of the search space.

3.2.2 Retrieval Aspect

Under normal operation, the search engine uses Lemur's vector space model techniques for retrieval, considering each document to be a vector of its constituent terms, each with some weight. This mode is referred to as "term-match", referring to the resulting effect of matching like documents based on term content. As an alternative to this, the search engine has a "category-match" mode. Lemur's vector space model routines are used here as well, but documents are considered instead to be vectors of their constituent MeSH category assignments, with weights. In order to represent the query in this format as well, it is labeled using the same process as for the search space, as described above. Searching then becomes a process of locating documents that have MeSH labelings like that of the query. This process occurs after any query translation, as a turnkey replacement for the traditional term-based monolingual retrieval.

Because category IDs are independent of the language of the documents to which they are applied, a category labeling can also be viewed as a mapping into a language-independent representation. This provides another avenue for translation: a query in one language, once transformed into a vector in category-space, may be used directly to retrieve documents from another language that are also in a category-space representation. The search engine supports this approach to translation in addition to the aforementioned approaches based on translating the query. In this case, PRF-based query expansion is still performed in term-space, before the query is transformed into a category-space representation. Pseudo-relevance feedback is still performed during retrieval in category-space, as well.

3.2.3 Retrieval Performance

Using solely "category-match" retrieval does not match the performance of traditional "term-match" retrieval. For monolingual cases, TREC average precision on the Springer dataset failed to exceed 64% of "term-match" performance. Scores for the assorted cross-lingual methods varied from about 40% of "term-match" performance to 72%.

To further investigate, "term-match" and "category-match" performance were compared using an alternate subset of the Springer documents for which human-assigned MeSH codes are available. Under this condition, "category-match" came closer to meeting "term-match" performance, achieving as much as 75% and 84% of the TREC average precision scores of "term-match" for the monolingual and cross-lingual cases, respectively. It has been determined that the performance of "term-match" retrieval could be improved upon with a

combination of both term-space and category-space representations, rather than using either alone (see section 4.2).

3.3 Corpus-Based Approaches (CMU, EIT, XRCE)

3.3.1 Similarity Thesauri (EIT, XRCE)

One purely corpus-based approach to CLIR is to build a similarity thesaurus (SimThes) over a parallel corpus. The similarity thesaurus contains words (adjectives, nouns, verbs) from the corpus, each accompanied by a set of words that appear in similar contexts and are thus similar in meaning. A similarity thesaurus can be built also over a monolingual corpus. It may then serve for query expansion in monolingual retrieval. In our case we built the similarity thesaurus over the parallel corpus. We were interested in German words and their similar counterparts in English. The similarity thesaurus is thus a bilingual lexicon with a broad translation set (in our case 10 similar English words per German word).

For example, for the German word *Myokardinfarkt* the similarity thesaurus contains the following 10 words in decreasing degrees of similarity:

infarction, acute myocardial infarction, myocardial, thrombolytic, acute, thrombolysis, crs, synchronisation, cardiogenic shock, ptca

Here we compare two different similarity thesaurus methods: EIT (Qui 1995) and XRCE (Gaussier et al. 2000). The main difference lies in the size of the context considered for retrieving translation equivalents: a pair of aligned sentences in the XRCE case, a pair of documents (or clusters of documents) in the EIT case. The lines DE2EN- XRCE -SimThes in table 9 have the results. The number in parentheses gives the number of similar terms used from the top of the similarity list. It is interesting to note that the number of relevant documents retrieved decreases if more than 10 similar words are used. Using between 5 and 10 similar documents thus seems like a good compromise between optimal precision and recall.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE2EN-SimThes (10)	0.2290	409	0.4492	0.3640
DE2EN-SimThes+all-comb.	0.2955	518	0.5761	0.4600
DE2EN-XRCE-SimThes (1)	0.3259	595	0.6910	0.6000
DE2EN-XRCE-SimThes (5)	0.3142	673	0.6763	0.5280
DE2EN-XRCE-SimThes (10)	0.2821	681	0.6064	0.4840
DE2EN-XRCE-SimThes (20)	0.2784	665	0.6049	0.4960

Table 5: CLIR results using a similarity thesaurus

3.3.2 Example-Based Thesaurus (CMU)

Another corpus-based approach to CLIR is the Example-based Thesaurus (EBT) method. EBT uses a sentence-aligned bilingual training corpus to find the terms that co-occur in context across languages, thus creating a corpus-based term-equivalence matrix. In this approach, terms are translated based on co-occurrence frequency in the context(s) defined by the document collection. Its results have proven superior to dictionary-based approaches (Yang et al 1998).

In order to create domain-specific or corpus-specific bilingual dictionaries automatically, we start from a large sentence-aligned bilingual corpus and generate a large thresholded term co-occurrence table (Brown 1997). This table is used as the dictionary for corpus-based (example-based) term substitution.

Co-occurrence dictionary generation is performed in two phases: First the co-occurrence matrix (indexed by source-language words on one axis and target-language words on the other) is generated. Each cell in the matrix represents the number of times the source-language word occurred in the same sentence pair as the target-language word. Then, given this matrix, we compute the conditional probability that if the term occurs in one language its counterpart (i.e. its candidate translation) also occurs in the other language within the same sentence pair, and vice-versa. If this probability is above a pre-set threshold **in both directions**, then the term translation is added into the dictionary. Should a term in one language co-occur with several terms in the other language with sufficient frequency to pass the conditional probability threshold, **all** are stored as candidate translations. This method has the nice property that adjusting the filtering thresholds allows us to tune a trade-off: stricter thresholds prevent spurious translations, but significantly reduce the possible translations; more lenient thresholds produce better yields, at the cost of allowing more spurious translations. Such corpus-based thesarus techniques are discussed in greater detail in (Brown 1997, Brown 1996).

3.3.3 Pseudo-Relevance Feedback (CMU)

Pseudo-relevance feedback (PRF), also known as “local feedback”, is a variation of the classic relevance feedback (RF) technique (Salton and Buckley 1990). Relevance feedback is a query expansion technique that adds terms in the **relevant** documents found in an initial retrieval to the query, and uses the expanded query for further retrieval. It typically improves performance in monolingual retrieval, compared to not using it. PRF differs from the true relevance feedback by assuming the top-ranking documents retrieved are all relevant. It is simpler because no user relevance judgments are required; but it is not always as effective as RF, because the top-ranking documents often include some irrelevant documents that may be misleading. Both positive and negative evidence was found in empirical studies with respect to the effect of PRF on retrieval accuracy (Hersh et al. 1994, Srinivasan 1996). We also found in a previous study (Yang et al 1998) that PRF cuts both ways, depending somewhat on how the queries were formulated originally.

Our primary interest in PRF has been to effectively cross the language barrier in translingual retrieval. Adapting PRF (and RF) to translingual retrieval is natural if a bilingual corpus is available (Carbonell et al 1997, Ballesteros and Croft 1997). That is, once the top-ranking documents are retrieved for a query in the source language, their translation mates (the corresponding documents in the target language) can be used to form the query in the target language.

The retrieval criterion in PRF for **monolingual** retrieval is defined to be:

$$\vec{q}' = \vec{q} + \sum_{i} \{\vec{d}_i | \vec{d}_i \in \text{kNN}(\vec{q})\}$$

$$sim(\vec{q}, \vec{d}) = \cos(\vec{q}', \vec{d})$$

where $\vec{q}$ is the original query, $\vec{q}'$ is the query after the expansion, $\text{kNN}(\vec{q})$ is the set of k Nearest Neighbors (most highly-ranked documents) retrieved using $\vec{q}$, and $i = 1, \ldots, k$.

Correspondingly, the retrieval criterion in PRF for **translingual** retrieval is defined to be:

$$\vec{q}_t = \sum_i \{\vec{g}_i | \vec{d}_i \in \text{kNN}(\vec{q}_s)\}$$

$$sim(\vec{q}_s, \vec{d}_t) = \cos(\vec{q}_t, \vec{d}_t)$$

where $\vec{q}_s$ is the query vector in the source language, $\vec{d}_i$ is the document vector in the source language and $\vec{g}_i$ is the document vector of its translation. $\vec{q}_t$ is the constructed query vector in the target language, and $\vec{d}_t$ is the target document in the search space. The length of each vector is m, the size of the term vocabulary after stemming and stop-word removal. Each element in the query and document vectors is weighted by $TF * IDF$.

4 Performance Evaluation (CMU, EIT)

In order to evaluate performance gain in information retrieval, several experiments have been carried out. The MuchMore document collection was used in combination with a query set and relevance assessments defined by medical experts from ZInfo.

Queries

The queries are short and usually consist of a complex noun phrase extended by attributes (including prepositional phrases) and co-ordination. Here are two examples. The complete list can be found in the appendix.

- *DE: Arthroskopische Behandlung bei Kreuzbandverletzungen.*
  EN: Arthroscopic treatment of cruciate ligament injuries.

- *DE: Indikation für einen implantierbaren Kardioverter-Defibrillator (ICD).*
  EN: Indication for implantable cardioverter defibrillator (ICD).

Relevance Assessments

For the experiments, we used relevance assessments based on 25 queries provided by the medical expert in the MUCHMORE project. We obtained relevance assessments based on the German documents as well as based on the English documents from two teams of experts. One team, which was organized by ZInfo in Germany consisted of medical professionals. The other team, which was led by CMU consisted of medical students. The two teams came up with two sets of relevant documents that were quite different: The ZInfo team finished with 959 relevant documents based on the German queries and documents. The CMU team defined 500 relevant documents for English. The main reason for this discrepancy is the different types of experts doing the assessments. The overlap was 382 documents while 118 were only deemed relevant by the CMU judges and 577 were only relevant for the ZInfo judges. In deliverable D9.1-2 we present a detailed list of numbers of relevant documents per query.

4.1 Concept-Based Methods and Similarity Thesaurus (EIT)

4.1.1 Retrieval System

For the retrieval experiments reported in this section, EIT used its commercial *relevancy* information retrieval system. In regular deployment this system extracts word tokens from

documents and queries alike and indexes them using a straight *lnu.ltn* weighting scheme. For the evaluation we adapted the *relevancy* system so that it indexes the information provided by the XML annotated documents and queries: word forms (tokens) and their base forms (lemmas) for all indexable parts-of-speech both for German and English. The indexable parts-of-speech encompass all content words, i.e. nouns (including proper names and foreign expressions), adjectives, and verbs (excluding auxiliary verbs).

4.1.2 Evaluation Measures

In subsequent tables we present the retrieval results in four columns: overall performance, measured as mean average precision (**mAvP**) - the mean of the precision scores after each relevant document retrieved; absolute **number of relevant documents retrieved**; average precision at 0.1 recall (**AvP01**); precision for the top 10 documents (**P10**).

4.1.3 Results

Vocabulary Overlap

A rough baseline for CLIR is using the tokens of the German queries directly for retrieval of the English documents. The idea is that the overlap in technical vocabulary between these languages will directly lead to some relevant documents. And indeed, this approach finds 66 relevant documents with German queries and English documents (cf. DE2EN-DE-token in table 6) and 86 relevant documents in the opposite direction. The best queries were those with the acronym *HIV* (which is the same in German and English) and with the Latin expression *diabetes mellitus*. For both these queries more than half the relevant documents were retrieved.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE2EN-DE-token	0.0512	66	0.1530	0.1160
EN2DE-EN-token	0.0504	86	0.1269	0.1480

Table 6: CLIR results via vocabulary overlap

It might be surprising that the overlap in technical vocabulary does not carry further than merely 66 or 86 out of 956 documents. But one must consider that often the roots of the technical terms are identical but the forms do not match because of differences in spelling and inflection (e.g. German *arthroskopische* vs. English *arthroscopic*). Stemming combined with some letter normalization (e.g. $k = c = z$) could lead to an increased recall, but has not been explored here.

Machine Translation of the Queries

As a second baseline we investigated the use of Machine Translation (MT) for translating the queries. We employed the PC-based system PersonalTranslator (linguatec, Munich) to automatically translate all queries from German to English. PersonalTranslator allows for restriction on the subject domain of the translation, for which the domains medicine and chemistry were selected here. Still, many words from the queries are not in its lexicon and remain untranslated (see the first example query below). Unfortunately the system does not segment compounds if it lacks knowledge of some of their parts. Therefore the word *Myokardinfarkts* is not segmented, although *Infarkt* is in the system's lexicon and could have been translated. Other queries are fully translated and almost perfect (see the second example query).

1. DE: *Behandlung des akuten Myokardinfarkts.*
   PT2001: *Treatment of the acute Myokardinfarkts.*
   EN: *Treatment of acute myocardial infarction.*

2. DE: *Möglichkeiten der Korrektur von Deformitäten in der Orthopädie.*
   PT2001: *Possibilities of the correction of deformities in orthopedics.*
   EN: *Approach of the correction of deformities in orthopedics.*

Many translations are incomplete or incorrect but still the automatically translated queries scored well with regard to recall. In table 7, line DE2EN-MT-PT2001, we see that these queries lead to 376 relevant documents at a (rather low) mean average precision of 0.1184.

In 2002 an improved version of PersonalTranslator was published. In line DE2EN-MTPT2002, we see that now the translated queries lead to an improved recall of 440 relevant documents at a still rather low mean average precision of 0.1381. In addition, linguatec provides a medical lexicon which is marketed as a separate product but which can be integrated into the MT system. This lexicon improves recall and precision significantly (see line DE2EN-MT-PT2002+MedLex). In fact it leads to one of the best results for German to English CLIR.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE2EN-MT-PT2001	0.1184	376	0.3382	0.2520
DE2EN-MT-PT2002	0.1381	440	0.3747	0.2920
DE2EN-MT-PT2002+MedLex	0.2393	543	0.5668	0.4440
EN2DE-MT-PT2001	0.0647	216	0.2150	0.1960
EN2DE-MT-PT2002	0.0618	212	0.1917	0.1800
EN2DE-MT-PT2002+MedLex	0.0723	215	0.2198	0.1840

Table 7: CLIR results: Queries automatically translated by PersonalTranslator

Surprisingly, this improvement does not apply to the opposite direction. When we search with English queries over German documents we started out with low precision and recall values with PersonalTranslator 2001 and they did not improve with the 2002 version nor with the medical lexicon. This runs counter to our observations that the translations did indeed get better with the new software.

EN: *New approach in cruciate ligament surgery*
PT2001: *Neuer Ansatz in einer cruciate Bandoperation*
PT2002: *Neuer Ansatz in einer cruciate Bandoperation*
PT2002+MedLex: *Neuer Ansatz in Kreuzbandeingriff*
DE: *Neue Erkenntnisse in der Kreuzbandchirurgie*

We believe that the results for English to German CLIR are so low because of the fact that the translation systems produces nice compounds (e.g. *Kreuzbandeingriff*), but these have to occur exactly as such in the documents. If they occur as separate words (e.g. *Kreuzband* and *Eingriff*) or in some other injected form (e.g. *Kreuzbandeingriffs*), the retrieval system will

not find them. If we want to use an MT system for translating English queries, it will be better to force such a system to avoid compounding.

Semantic Annotation

Now let us compare these results with the semantic codes annotated in the documents and queries. This means we are using the semantic annotation of the German queries to match the semantic annotation of the English documents. One could say that we are now using the semantic annotation as an interlingua or intermediate representation to bridge the gap between German and English.

Table 8 has the results. This time the UMLS terms lead to the best results with respect to recall, but MeSH is (slightly) superior regarding precision. EuroWordNet leads to the worst precision and the semantic relations have only a minor impact due to their specificity. If we combine all semantic information, we achieve the best recall (404) and mean average precision (0.1774).

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE2EN-EWN	0.0090	111	0.0311	0.0160
DE2EN-UMLS	0.1620	366	0.3724	0.2800
DE2EN-MeSH	0.1699	304	0.3888	0.2600
DE2EN-Semrel	0.0229	23	0.0657	0.0480
DE2EN-all-combined	0.1774	404	0.3872	0.2720

Table 8: CLIR results: using semantic codes

Summary

In cross-language retrieval machine translation of the queries fared surprisingly good for German to English retrieval, especially if supplemented with a domain-specific bilingual lexicon (i.e. a medical lexicon). Machine Translation is rather bad for English to German because of the compound word problem. Semantic annotation using UMLS and MeSH reaches a level comparable to Machine Translation without the domain-specific lexicon. Semantic annotation is however superseded by a similarity thesaurus built over the parallel corpus. When using a similarity thesaurus built on document alignment, the highest overall performance resulted from a combination of this similarity thesaurus with the semantic information.

4.2 MeSH Concept-classification and Corpus-Based Methods (CMU)

The evaluation numbers presented in Table 9 below represent the TREC average precision performance ("**mAvP**") on the test-set half of the MuchMore Springer dataset using only the ZInfo-generated German relevance judgements, as decided at the Hvar workshop. Both our PRF and EBT engines were trained on the training-set half of the Springer dataset, which was also used for query expansion. Since concept-based approaches require MeSH labels and the Springer dataset has none, we used the OHSUMED-87 corpus to train our concept-based approaches. OHSUMED-87 is English-only, so, in order to obtain labelled German training data, we used kNN trained on OHSUMED-87 to label the English half of the Springer training set, and transferred those labels to the German half.

In all cases, the numbers in the table represent for each case the best method in that condition. For "Terms Only" and "Terms+Concepts/German to English" this was EBT; for "Concepts Only" and "Terms+Concepts/English to German" this was PRF.

Both our PRF and EBT methods are competitive, with PRF doing better English to German and EBT better for German to English. The concept-based approaches alone did not do nearly as well as these traditional term-based approaches. Concept-based performance when retrieving German documents is particularly poor, which is partially attributable to the fact that OHSUMED-87 is English-only: whereas we were able to label the English half of the Springer test set directly using OHSUMED-87, labelling the German half used the German half of the Springer training set, thus suffering two levels of machine assignment.

However, as shown in Table 9, the **combination** of term-based and concept-based approaches produced an improvement in all cases except English monolingual. We believe that this was probably due to overfitting to our training data during parameter tuning.

	Terms Only	Concepts Only	Terms+ Concepts	% Improvement
English to English	0.57	0.46	0.55	-3.51%
German to German	0.34	0.24	0.39	14.71%
English to German	0.53	0.47	0.58	9.43%
German to English	0.32	0.19	0.33	3.13%

Table 9: Comparison of CMU's best traditional (term-based) and concept-based approaches. All scores are TREC average precision ("mAvP" in the tables in previous sections.)

5 User Evaluation[19] (ZInfo)

The MuchMore project identified several major needs where language constitutes a relative barrier in medical information use and retrieval:

- Reducing the gap between medical documentation and multilingual media
- Eliminating the language barrier between media and medical professionals
- International comparison and benchmarking of medical records
- Multilingual medical software solutions

A usability test of the MuchMore prototype along these lines was carried out by ZInfo with physicians from different hospitals and with different specializations, which participated in an advanced course on medical informatics. The information access test case was the connection between a patient's medical history and the medical literature. The gain of the technologies and resources developed in the project was measured against "traditional" existing methods. The expected results were information about the usability and acceptance of different aspects of the developed prototype in a realistic scenario, and about required improvements needed for the construction and deployment of an information access system based on the technologies and resources developed in the project.

5.1 Evaluation Group and Measures

All 10 users that participated in the evaluation spoke German and English fluently. Other spoken languages included French, Russian and Romanian either as a mother language or as a

[19] For a detailed account of the user evaluation, please refer to deliverable D10.1

second language learned in school. 5 users were residents and the other 5 were specialists taking their MI post-graduate specialization course. 5 worked in a university clinic and 5 in a non-educational clinic. Represented specialities included: 3 in surgery, 2 in internal medicine, 1 general practitioner, 2 in psychiatry, 1 radiologist and 1 in hygiene and environmental medicine. All had over two years of Internet IR experience, the majority using Internet for literature retrieval once a week. Used medical Internet engines were represented by Medline and PubMed[20] for English documents and DIMDI and Dr. Antonius[21] for German documents.

Each judge provided 3 anonymous bilingual query texts (case reports, patient discharge letters, epicrisis) as input. In all there were 60 retrieval-runs - 30 for German and 30 for English documents. User profile and result forms for each query text were then completed.
Some examined profile attributes were:

1. Spoken languages
2. Medical degree and institution type (hospital, practice, university-hospital etc.)
3. Speciality
4. Bibliographic resource preferences and timely needs
5. Internet experience in document retrieval

After completion, the inquiry results were grouped and a statistical analysis was performed identifying several usability, relevance and general acceptance parameters for evaluating the prototype:

1. Relevance of extracted terms and relationships for the patient document
2. English/German document ratio in the first 20 retrieval results.
3. Document relevance in the first 20 results.
4. Machine versus human document ranking of the most relevant English/German document
5. Relevance of extracted terms and relationships from the most relevant English/German document for a future iterative search.
6. Subjective estimation of the prototypes' "translation through concepts" feature.
7. General acceptance and usability estimation.

5.2 Results

Term and Relationship Extraction for Query building

The user relevance of the extracted terms and relationships for a given profile and clinical case was examined taking into account the language of the input document:

	Concepts	**Relations**
German Input Document	32,82%	23,18%
English Input Document	16,87%	11,56%

[20] http://www.pubmedcentral.nih.gov/
[21] http://www.dr-antonius.de/

After retrieval and identification of the most relevant English/German document in the first 20 results, terms and relationships were also extracted in order to evaluate the relevance of the most relevant retrieved document in an iterative search:

	Most relevant German Abstract		Most relevant English Abstract	
	Concepts	Relations	Concepts	Relations
German Input Document	24,66%	18,07%	18,02%	18,81%
English Input Document	24,53%	19,59%	17,31%	16,93%

On a scale from 1 to 5 the user was asked to subjectively estimate the accurateness and usability of extracted terms and relationships for his/her profile and according to the medical case:

	Concepts	Relations
German Input Document	2,83	3,50
English Input Document	3,03	3,50
Total	2,92	3,21

The subjective estimation roughly follows the concept and relationship extraction metrics.

German/English partition ratio

We also examined the German/English partition when retrieving documents from the Springer collection. Independent of the language of the input document the ratio was always in favour of the English language although, as we have seen the choice of English concepts for query building/expansion was significantly lower as that of German concepts:

	Retrieved German Documents	Retrieved English Documents	G/E partition
German Input Document	1236	1393	47% / 53 %
English Input Document	887	1288	41% / 59%

Cross-lingual precision

The cross-lingual precision of the top 20 results (for 600 retrieved documents) dependent upon the language of the input document can be observed:

Input Document Language	Relevant German Documents	Relevant English Documents	German Document Precision	English Document Precision
German	58	61	9,67%	10,17%
English	37	62	6,17%	10,33%
Total	95	123	7,92%	10,25%

Search-engine versus expert ranking

Interestingly enough the results were also compared over search-engine relevance ranking versus expert ranking and English most relevant documents found were definitely better placed than German ones although the difference was only slight.

	most relevant German document rank	most relevant English document rank	Distance from 1st - German	Distance from 1st - English
German Input Document	7,3	6,52	5,3	4,91
English Input Document	7,9	7,22	5,8	5,57

User CLIR satisfaction

Finally, the judges gave an overall verdict of their satisfaction with the CLIR features of the system on the one hand and the general satisfaction with the MUCHMORE search methodology. CLIR satisfaction was constant independent of direction. The users subjectively seem to be more satisfied with MUCHMORE handling English Input documents as with documents in their mother language:

	User Satisfaction
German to English	2,43
English to German	2,57
CLIR Satisfaction	2,44

Overall user satisfaction

On a scale from 1 to 5 users were asked to estimate the overall satisfaction and usability of the language technologies implemented in the MM prototype:

MUCHMORE	User Satisfaction
German Input	3,43
English Input	3,13
MM Prototype satisfaction	3,2

6 Conclusions

The MuchMore project produced a prototype for medical cross-lingual information retrieval that integrates a number of hybrid approaches, combining corpus-based (use of statistics and machine-learning) and concept-based (use of knowledge bases) methods. These approaches include semantic annotation and automatic thesaurus construction.

Semantic annotation in the MuchMore approach is based on linguistic pre-processing (PoS-tagging and morphological analysis) and includes sense disambiguation to map ambiguous terms to the most appropriate concept, next to term and relation extraction to extend existing knowledge bases upon need. Semantic annotation (of UMLS terms) showed a positive effect on precision as well as recall in the CLIR task. Sense disambiguation (of UMLS terms) showed some effect on mean average precision, but not in the high precision area. Semantic annotation and disambiguation of EuroWordNet terms showed no positive effects.

Extraction of bilingual term pairs from parallel as well as comparable corpora showed good results. Most significant in term extraction from comparable corpora is the use of existing knowledge bases (in this case UMLS/MeSH). CLIR experiments with the extracted term pairs showed an increase in recall and precision, although precision drops if more than 5 (top) translation candidates are included as term pairs.

Relation extraction remains a very difficult topic of research, which implies a rather intensive involvement of domain experts in development and evaluation. In experiments with filtering (i.e. disambiguation) of extracted relations no positive effect could be shown on the CLIR task. On the other hand, extraction of novel instances for existing relations showed an increase in both precision and recall.

Automatic thesaurus construction methods (i.e. similarity thesaurus, EBT) showed consistently very good results in the CLIR task. Nevertheless, best results were obtained in combining these with existing (manually constructed) knowledge bases (i.e. UMLS/MeSH).

From the user point of view, medical experts were interested in the different functionalities provided by the MuchMore prototype: query construction tool, meta-search engine and summarization tool. The query construction tool provides medical professionals with a novel functionality to interactively construct a query from an uploaded electronic patient record. The meta-search engine allows the user to combine different methods in a transparent way, i.e. they can try out different settings to obtain the best result without knowledge of the underlying methods. The summarization tool provides a useful option to automatically summarize all or a selection of returned documents.

References

Lisa Ballesteros and Bruce Croft *Phrasal translation and query expansion techniques for cross-language information retrieval*. 20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), 1997, pp. 85-91.

Ralf D. Brown *Example-Based Machine Translation in the Pangloss System*. Proceedings of the Sixteenth International Conference on Computation Linguistics, pp. 169-174, 1996.

Ralf D. Brown *Automated Dictionary Extraction for 'Knowledge-Free' Example-Based Translation*. Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation, 1997.

J. G. Carbonell, and J. Goldstein *The use of MMR, diversity-based reranking for reordering documents and producing summaries*. In Proceedings, 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 335-336, Melbourne, Australia, August 1998.

J. G. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee *Translingual Information Retrieval: A Comparative Evaluation*. Proceedings of IJCAI-97, Nagoya, Japan, August 1997. Distinguished paper award.

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

Eric Gaussier, David Hull, and Salah Ait-Mokhtar. *Term alignment in use: Machine-aided human translation*. In Jean Veronis, editor, *Parallel Text Processing*. Kluwer, Dordrecht, 2000.

W. Hersh, C. Buckley, T.J. Leone, and D. Hickman *OHSUMED: An Interactive Retrieval Evaluation and New Large Text Collection for Research*. 17th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94), pp. 192-201, 1994.

O. Plaehn, Brants Th. *Annotate – An Efficient Interactive Annotation Tool*. In: Proceedings of the 6th Conference on Applied Natural Language Processing ANLP, Seattle, WA, 2000.

G. Salton and C. Buckley *Improving Retrieval Performance by Relevance Feedback*. Journal of American Society for Information Sciences, 1990, vol. 41, pp. 288-297.

P. Srinivasan *Optimal Document-Indexing Vocabulary for MEDLINE*. Information Processing & Management, 32(5), pp. 503-514, 1996.

Y. Qui. 1995. *Automatic Query Expansion Based on a Similarity Thesaurus*. Phd thesis, ETH Zurich.

S. Vintar, P. Buitelaar, B. Ripplinger, B. Sacaleanu, D. Raileanu, and D. Prescher. *An efficient and flexible format for linguistic and semantic annotation*. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, Spain, May 29-31, 2002.

Vossen, P. 1997. *EuroWordNet: a multilingual database for information retrieval.* In: Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.

Y. Yang, J.G. Carbonell, R.E. Frederking, and R. Brown *Translingual information retrieval: learning from bilingual corpora*. Artificial Intelligence Journal special issue: Best of IJCAI-97 (invited submission), 1998.

Appendix A: Dissemination and Concertation

Demos

The MuchMore project produced a number of demonstration systems to display various results in research and development on cross-lingual information retrieval and multilingual domain modelling (i.e. term clustering for sense and semantic relation discovery):

MuchMore Cross-Lingual Information Retrieval Prototype: Meta-Search Engine, Query Construction Tool and Multi-Document Summarization

Search Engine @ CMU (incl. Multi-Document Summarization Demo)

Search Engine @ Stanford University, CSLI

Search Engine @ Eurospider Information Technologies AG

DFKI MuchMore Annotation Tool + Annotation Display Tool MMV

Stanford University, CSLI Term Clustering (Sense Discovery) Demo

DFKI Term Clustering (Semantic Relation Discovery) Demo – in cooperation with the Jozef Stefan Institute, Ljubljana, Slovenia

These demos are available under:

http://muchmore.dfki.de/demos.htm

Publications

All MuchMore related publications are available under: http://muchmore.dfki.de/pub.html

2001

- Paul Buitelaar and Bogdan Sacaleanu *Ranking and Selecting Synsets by Domain Relevance* In: Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. Carnegie Mellon University, Pittsburgh, 3 and 4 June 2001.
- Paul Buitelaar *The SENSEVAL-II Panel on Domains, Topics and Senses* To Appear In: Proceedings of SENSEVAL-II, Toulouse, France, July 2001.
- Paul Buitelaar, Jan Alexandersson, Tilman Jaeger, Stephan Lesch, Norbert Pfleger, Diana Raileanu, Tanja von den Berg, Kerstin Klöckner, Holger Neis, Hubert Schlarb *An Unsupervised Semantic Tagger Applied to German* In: Proceedings of Recent Advances in NLP (RANLP), Tzigov Chark, Bulgaria, 5-7 September, 2001.
- Detlef Prescher *Novel Properties and Well-Tried Performance of EM-Based Multivariate Clustering* In: Proceedings of the EuroConference on Recent Advances in Natural Language Processing (RANLP-01). Tzigov Chark, Bulgaria, September, 2001.

- Detlef Prescher *Inside-Outside Estimation Meets Dynamic EM* In: Proceedings of the 7th International Workshop on Parsing Technologies (IWPT-01). Beijing, China, October 2001.
- Monica Rogati, Yiming Yang. Cross-Lingual Pseudo-Relevance Feedback Using a Comparable Corpus. CLEF 2001: 151-157

2002

- Paul Buitelaar, Bogdan Sacaleanu *Extending Synsets with Medical Terms* In: Proceedings of the First International WordNet Conference, Mysore, India, January 21-25, 2002.
- Ralf D. Brown, "Corpus-Driven Splitting of Compound Words", In Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002). Keihanna, Japan, March 2002.
- Bärbel Ripplinger, Špela Vintar, Paul Buitelaar *Cross-Lingual Information Retrieval through Semantic Annotation*. In: Proceedings of NLPBA2002, Cyprus, March 8-10, 2002.
- Špela Vintar, Paul Buitelaar, Bärbel Ripplinger, Bogdan Sacaleanu, Diana Raileanu, Detlef Prescher *An Efficient and Flexible Format for Linguistic and Semantic Annotation*. In: Proceedings of LREC2002, Las Palmas, Canary Islands - Spain, May 29-31, 2002.
- Diana Raileanu, Paul Buitelaar, Jörg Bay, Špela Vintar *An Evaluation Corpus for Sense Disambiguation in the Medical Domain*. In: Proceedings of LREC2002, Las Palmas, Canary Islands - Spain, May 29-31, 2002.
- Dominic Widdows, Beate Dorow, and Chiu-Ki Chan. *Using Parallel Corpora to enrich Multilingual Lexical Resources*. Third International Conference on Language Resources and Evaluation: LREC2002, Las Palmas, May 2002, pages 240-245.
- Dominic Widdows and Beate Dorow. *A Graph Model for Unsupervised Lexical Acquisition.* 19th International Conference on Computational Linguistics, Taipei, August 2002, pages 1093-1099.
- Dominic Widdows, Beate Dorow, and Scott Cederberg. *Visualisation Techniques for Analysing Meaning.* Fifth International Conference on Text, Speech and Dialogue, Brno, Czech Republic, September 2002, pages 107-115.
- Martin Volk, Paul Buitelaar *A Systematic Evaluation of Concept-based Cross Language Information Retrieval in the Medical Domain* In: Proceedings of DIR-2002: 3d Dutch-Belgian Information Retrieval Workshop, Leuven, Belgium, December 6th, 2002.
- Volk, Martin / Ripplinger, Bärbel / Vintar, Spela / Buitelaar, Paul / Raileanu, Diana / Sacaleanu, Bogdan: *Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval*. In: International Journal of Medical Informatics, Volume 67:1-3, Dec. 2002.
- Jean-Michel Renders, Hervé Déjean, Éric Gaussier. *Assessing Automatically Extracted Bilingual Lexicons for CLIR in Vertical Domains: XRCE participation to the GIRT task of CLEF 2002*. In: Cross-Language Information Retrieval and Evaluation, Lecture Notes in Computer Science. Springer-Verlag.

2003

- Buitelaar, Paul / Declerck, Thierry: *Linguistic Annotation for the Semantic Web*. In: Siegfried Handschuh, Steffen Staab (eds.) Annotation for the Semantic Web, IOS Press, January, 2003.
- Buitelaar, Paul / Declerck, Thierry / Sacaleanu, Bogdan / Vintar, Spela / Raileanu, Diana / Crispi, Claudia: *A Multi-Layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations*. In: Proceedings of EACL 2003 Workshop on Language Technology and the Semantic Web (NLPXML'03), Budapest, Hungary, April 2003.
- Sacaleanu, Bogdan / Volk, Martin / Buitelaar, Paul: *A Cross-Language Document Retrieval System Based on Semantic Annotation*. In: Proceedings of EACL 2003 Demo Session, Budapest, Hungary, April 2003.
- Volk, Martin / Vintar, Špela / Buitelaar, Paul: *Ontologies in Cross-Language Information Retrieval*. In: Proceedings of WOW2003 (Workshop Ontologie-basiertes Wissensmanagement), Luzern, Switzerland, April 2003.
- Dominic Widdows. *Unsupervised methods for developing taxonomies by combining syntactic and statistical information*. In Proceedings of HLT/NAACL 2003, Edmonton, Canada, June 2003, pages 276-283.
- Widdows, Dominic / Peters, Stanley / Cederberg, Scott / Chan, Chiu-Ki / Steffen, Diana / Buitelaar, Paul: *Unsupervised Monolingual and Bilingual Word-Sense Disambiguation of Medical Documents using UMLS*. In: Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine, Sapporo, Japan, July 11th, 2003.
- Špela Vintar, Paul Buitelaar, Martin Volk *Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval* To Appear In: Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM), September 22nd, 2003, Cavtat-Dubrovnik (Croatia).
- Špela Vintar, Ljupčo Todorovski, Daniel Sonntag, Paul Buitelaar *Evaluating Context Features for Medical Relation Mining* To Appear In: Proceedings of the ECML/PKDD Workshop on Data Mining and Text Mining for Bioinformatics, September 22nd, 2003, Cavtat-Dubrovnik (Croatia).

Presentations

The project was presented at:

- Panel on Cross-Lingual Information Retrieval at the ASIS 2000 Conference (American Society for Information Science and Technology), Chicago, USA, 2000
- First CLASSiks workshop (Madrid, Spain, 4th/5th May, 2001) on concertation efforts in the area of Cross-Lingual Information and Knowledge Management. The main objectives of the project were presented and possibilities for closer cooperation were investigated. More specifically, we discussed with representatives of the LIQUID project on how to combine efforts in user and technical evaluation.
- Second CLASSiks workshop (Schloss Dagstuhl, Germany, October 1st/2nd, 2001) on concertation efforts in the area of Cross-Lingual Information and Knowledge Management.
- OntoWeb SIG5 meeting on "Language Technology in Ontology Development and Use" at the OntoWeb3 meeting of the EU-IST Thematic Network OntoWeb

(http://www.ontoweb.org), June 13th/14th, 2002, Sardinia, Italy: http://ontoweb-lt.dfki.de/events.htm
- Panel on the Semantic Web: A New Challenge for Language Technology? COLING-02, Taipei, August 30th, 2002
- Final CLASSiks workshop (Berlin, Germany, September 28th, 2002) on concertation efforts in the area of Cross-Lingual Information and Knowledge Management.
- Institute for InfoComm Research, January 2003, Singapore – in the context of the Semantic Web seminar series
- Institut de Lingüística Aplicada (IULA), Universitat Pompeu Fabra, February 2003, Barcelona, Spain -- in the context of the PhD program in Applied Linguistics (NLP area)
- CLIF (Computational Linguistics in Flanders), April 25th 2003, Brussel, Belgium

The project organized an international workshop with a select group of invited experts in medical information access, cross-lingual information retrieval, and semantic annotation. The purpose of the workshop was to disseminate project results and obtain expert feedback from researchers and potential users. Therefore, the experts invited included researchers and developers that covered at least one of the following requirements:

- highly experienced in their research area (i.e. medical information access, cross-lingual information retrieval, semantic annotation)
- up to date with current technology available in medical information access
- knowledgable representatives of relevant user groups

Invited experts were:

- Wilhelm Gaus (Biometrics and Medical Documentation, Medizinische Fakultät, Ulm University, Germany)
- Julio Gonzalo (NLP Group, UNED, Madrid, Spain)
- Greg Grefenstette (Clairvoyance Corporation, USA/France)
- Liz Liddy (Center for NLP, School of Information Studies, Syracuse University, USA)
- Stuart Nelson (Medical Subject Headings Section, National Library of Medicine, NIH, USA)
- Alan Rector (Medical Informatics Group, Computer Science Department, University of Manchester, UK)
- Michael Schopen (German Institute for Medical Documentation and Information, Cologne, Germany)
- Hinrich Schütze (Novation BioSciences Inc., USA)

Industrial Awareness

The project was presented to representatives of:

- SAP AG, Germany: June 28th, 2001
- iAS AG, Marburg, Germany: June 25th, 2001
- Siemens Medical Solutions, Germany: September 12th, 2001
- Language and Computing NV, Zonnegem, Belgium: December 5th, 2001
- Software Research Center, Ricoh Co., Ltd., Japan: March 21st, 2002

- Brockhaus Duden Neue Medien GmbH, Mannheim, Germany: September 5th, 2002
- SmartBotTechnologies GmbH, Bad Soden a. Ts., Germany: November 27th, 2002
- ID GmbH (active in medical documentation), Berlin, Germany: November 2002

Appendix B: Total Project Effort in PM

Total Project Effort (person-months)							
Participant's short name / **Workpackage**	**DFKI**	**XRCE**	**ZInfo**	**EIT**	**CMU**	**CSLI**	**TOTAL**
WP0	27						**27**
WP1	3	2	2	1,7	2	2	**12,7**
WP2	3		8				**11**
WP3	7			5,3	6	6	**24,3**
WP4.1	12	0,3	3,2	2,5	2		**20**
WP4.2	10,2		4	0,3			**14,5**
WP4.3			12			8	**20**
WP5	25		5,2			32	**62,2**
WP6					3		**3**
WP7.1		21,3		5,6	4		**30,9**
WP7.2	18,5		3,4			10	**31,9**
WP7.3	5,5		1				**6,5**
WP8.1	1,5	2		7,9	7	9	**27,4**
WP8.2	1,5			2	4	2	**9,5**
WP8.3	6			4,4		3	**13,4**
WP9.1			1	16,7	8		**25,7**
WP9.2				8,6			**8,6**
WP10	0,5		4,8				**5,3**
TOTAL	**120,7**	**25,6**	**44,6**	**55**	**36**	**72**	**353,9**