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 24*
Actual date of delivery	*Month 36*
Deliverable number	*D9.1_2*
Deliverable title	*Technical Evaluation Report*
Type	*Report*
Status & version	*Final Version*
Number of pages	*30*
WP contributing to the deliverable	*WP9.1, WP9.2*
WP / Task responsible	*CMU*
Author(s)	*Bryan Kisiel (CMU), Martin Volk (EIT)*
EC Project Officer	*Yves Paternoster*
Keywords	*Cross-Lingual Information Retrieval; Evaluation; Corpus-based Approaches; Concept-based Approaches*
Abstract (for dissemination)	*The experiments reported in this paper were performed within the framework of the MUCHMORE project, which aimed at systematically comparing several concept-based and corpus-based methods in cross-language medical information retrieval. Primary goals of the project included: Developing and evaluating methods for the effective use of multilingual thesauri for the purpose of Cross-Lingual Information Retrieval, including the linguistic and semantic annotation of English and German medical texts; Subsequently evaluating and comparing the impact of this information on retrieval.*

Introduction

The experiments reported in this paper were performed within the framework of the MUCHMORE project, which aims at systematically comparing several concept-based and corpus-based methods in cross-language medical information retrieval. Primary goals of the project included:

1. Developing and evaluating methods for the effective use of multilingual thesauri for the purpose of Cross-Language Information Retrieval (CLIR), including the semantic annotation of English and German medical texts;

2. Subsequently evaluating and comparing the impact of such semantic information for this purpose.

The report is organized as follows. In section 1 we present the evaluation work carried out by the MUCHMORE partner EIT. The focus of this work is on evaluating and comparing the effect of linguistic and semantic annotation (as described in deliverable D4.1) on the CLIR task. In section 2 we present the evaluation work carried out by the MUCHMORE partner CMU. Here, the focus is on evaluating and comparing the concept-based classification approach to CLIR in combination with different corpus-based methods. Finally, in section 3 we present the conclusions of these studies.

1 EIT Evaluation

This section of the report presents first in subsection 1.1 the resources and approaches for linguistic and semantic annotation used. In subsection 1.2 we describe the retrieval experiments and their results using different indexing features.

1.1 Annotation

The main document collection used in the MUCHMORE[1] project is a parallel corpus of English-German scientific medical abstracts obtained from the Springer LINK web site[2]. The corpus consists of approximately 9000[3] documents with a total of one million tokens for each language. Abstracts are taken from 41 medical journals (e.g. *Der Nervenarzt*, *Der Radiologe*, etc.), each of which constitutes a homogeneous medical sub-domain (e.g.

[1] MUCHMORE was sponsored by the European Union under grant IST-1999-11438. The European project partners are DFKI GmbH, LT Department, Saarbr¨ucken; XEROX Research Centre Europe, Grenoble; ZInfo, Klinikum der J.W. Goethe-Universit¨at Frankfurt; and Eurospider Information Technology AG, Zurich. The project also includes two partners in the US: Carnegie Mellon University, LT Institute, and Stanford University, CSLI. For details see http://muchmore.dfki.de

[2] http://link.springer.de

[3] After the evaluation runs were finished we realized that due to a misunderstanding only around 7800 documents per language were annotated and used in the experiments. This means that with respect to the ZInfo relevance assessments around 180 documents were not in the collection, and with respect to the CMU relevance assessments around 85 documents were missing. Since the documents were missing in all of our evaluation experiments, the relative results (method A compared to method B) are correct. But the absolute values, in particular the recall values, are clearly worse than they would have been if given the complete document collection.

Neurology, Radiology, etc.). Corpus preparation and annotation was done by DFKI. It included removing special tags and symbols in order to produce a clean, plain text version of each abstract,consisting of a title, text and keywords. The corpus was then linguistically annotated using standard tools for shallow processing: a tokenizer, a statistical part-of-speech tagger, a morphological analyser and a chunker for phrase recognition.

1.1.1 The Annotation Resources

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

In the MUCHMORE project we assigned 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 ending semantic relations. Appropriate EuroWordNet synset codes were assigned if a word or an expression belonged to a EuroWordNet synset [Buitelaar and Sacaleanu 2001].

We strictly apply semantic annotation in a monolingual way, in which information available from parallel documents is not considered. This is to ensure that the approach

will be applicable to any multilingual document collection (for our purposes here in English and German) and not only to parallel document collections.

1.1.2 The XML Annotation Format

Both morpho-syntactic (part-of-speech, morphology, phrases) and semantic (terms, semantic relations) annotation are integrated in a multi-layered XML annotation format, which organizes various levels as separate tracks with options of reference between them via indices. The 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>
   <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>
```

The linguistic analyzer 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) synsets. For the example sentence we determined that word w21, *perception*, has four EWN senses, related to *perceiving - sensing, perception*, and *perceptual experience*. We have also experimented with word sense disambiguation methods to 4 cut down on ambiguities concerning EWN senses, based on methods described in [Buitelaar and Sacaleanu 2001]. Evaluation of the disambiguation module is undertaken as part of the CLIR evaluation task (comparing disambiguated and non-disambiguated versions of the annotated document collection), as well as separately by using a manually tagged lexical sample corpus [Raileanu et al. 2002].

```
<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”. And 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 of our semantic information are the semantic relations that we derive from the UMLS Semantic Network. This network 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"/>
```

1.2 Evaluation

1.2.1 Set of Queries and Relevance Assessments

In order to evaluate whether the semantic annotations result in a performance gain in information retrieval, several experiments have been carried out. We used our document collection (the set of medical abstracts described in section 1.1) as well as a query set defined by medical experts. The OSHUMED collection would not have been appropriate for the MUCHMORE project due to its monolingual nature (documents and queries are only available in English).

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 the appendix we present a detailed list of numbers of relevant documents per query.

Because the MUCHMORE corpus is parallel, we decided to use the ZInfo relevance assessments for most of our experiments in order to get comparable data. In these assessments the number of relevant documents per query varies between 7 and 104. In section 1.2.4.5 we will compare the evaluation results for the two sets of relevance assessments.

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).

1.2.2 The Retrieval System

For the retrieval experiments we used the commercial *relevancy* information retrieval system from Eurospider Information Technology AG. 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 theoretical background of this scheme see [Schäuble

1997]). In addition the system can index word stems derived from a lexicon-based (Celex) stemmer for German, and a Porter-like stemmer for English [Wechsler et al. 1997]. This stemming was not used in the evaluation experiments reported here.

For the MUCHMORE evaluation runs 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).

- Indexable PoS tags for German (from the STTS POS-tag set[4]): ADJA, ADJD, FM, NN, NE, TRUNC, VVFIN, VVIMP, VVINF, VVIZU, VVPP
- Indexable PoS tags for English: FW, JJ, JJR, JJS, NN, NNP, NNPS, NNS, VB, VBD, VBG, VBN, VBP, VBZ

The decision whether a word is being indexed thus depends on the automatically assigned PoS-tag. Tagging errors within the content-bearing word classes do not matter (e.g. the confusion of a noun with an adjective). But if a content word is tagged as a function word (e.g. an adjective erroneously tagged as adverb), this word will be missing in the index.

In addition, all semantic information was indexed in separate categories each: EuroWord-Net terms, UMLS terms, semantic relations, and MeSH terms.

1.2.3 The Evaluation Measures

In all subsequent tables we present the retrieval results in four columns. The first column contains the overall performance, measured as mean average precision (**mAvP**) as has become customary in the TREC experiments (cf. [Gaussier et al. 1998])[5]. This figure is computed as the mean of the precision scores after each relevant document retrieved. The value for the complete evaluation run (i.e. the set of all queries) is the mean over all the individual mean average precision scores. This value contains both precision and recall oriented aspects and is the most commonly used summary measure. In the second column we present the absolute **number of relevant documents retrieved**, a pure recall measure. Third, we present the average precision at 0.1 recall (**AvP01**). This can be interpreted as answering the question: How many documents do I have to browse through from the top of the list until I reach 10% recall?

According to [Eichmann et al. 1998], the effectiveness within the high precision area is measured assuming that users are most interested to get relevant documents ranked topmost in the result list. Because this number can vary substantially for different queries, we consider also the precision figures for the topmost documents retrieved (in column four). There we focus on the top 10 documents (**P10**).

[4] http://www.coli.uni-sb.de/sfb378/negra-corpus/stts.asc
[5] The evaluation measures are described in detail in:
http://trec.nist.gov/pubs/trec10/appendices/measures.pdf

1.2.4 Monolingual Evaluation Runs

MUCHMORE aims first and foremost at cross-language retrieval (CLIR). In order to assess the CLIR performance, monolingual experiments in German and English were conducted acting as baselines for the cross-language experiments[6]. In the monolingual experiments the queries and the documents are of the same language. Most CLIR systems achieve only up to 75% precision compared to monolingual IR (cf. [Schäuble and Sheridan 1998]), and one goal of MUCHMORE is to check whether semantic annotation improves CLIR performance.

For each language, we produced a baseline performance by indexing only the tokens in both the documents and the queries. We call these baselines DE-token and EN-token. Some recent works have shown, that at least for German a linguistic-based stemming and decompounding is beneficial for retrieval, and therefore two evaluation runs based on linguistic stemming were produced, which we termed DE-token-lemma and EN-token-lemma. In table 1 we present the results of the monolingual German retrieval experiments.

1.2.4.1 German Monolingual Retrieval

In the baseline experiment for German (DE-token) the system finds only 322 relevant documents (out of 956; cf. table 1). The mean average precision is thus low (mAvP = 0.16), but the average precision in the top ranks is acceptable (AvP = 0.56). So, the few documents that are found are often ranked at the top of the list. On average there are 4.16 relevant documents among the 10 top ranked documents. This is expressed by the value of 0.4160 for P10.

The importance of good linguistic stemming and decompounding is shown by the second experiment (DE-token-lemma), which achieves a recall gain of 60% compared to DE-token.

In parallel, the precision figures have improved substantially. Lemmatization was done with a general-purpose morphological analyzer (as described in [Volk et al. 2002]). In section 1.2.4.3 we will explain how some heuristic morphology rules improve the lemmatization step.

The impact of the different types of semantic information was determined one by one, but always in combination with tokens and lemmas. We wanted to support the hypothesis that semantic information will improve the precision over pure token and lemma information. It turns out that the MeSH codes are the most useful indexing features whereas the EuroWordNet terms (EWN), without disambiguation in our current experiments, are the worst. Using MeSH codes slightly increases recall (from 516 to 526) but most impressively improves average precision (from 0.2180 to 0.2452). The positive impact of the UMLS terms is less visible and - as was to be expected - the very specific semantic relations (Semrel) have hardly any impact. Using the EuroWordNet terms in this combination with lemmas and tokens degrades the overall performance.

[6] Some of the evaluation results have been published in [Volk et al. 2002] and [Volk and Buitelaar 2002].

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE-token	0.1600	322	0.5622	0.4160
DE-token-lemma	0.2180	516	0.5967	0.4720
DE-token-lemma-EWN	0.1980	500	0.5571	0.4520
DE-token-lemma-UMLS	0.2236	509	0.5895	0.4640
DE-token-lemma-MeSH	0.2452	526	0.6356	0.5120
DE-token-lemma-Semrel	0.2224	516	0.5841	0.4640

Table 1: Results of the monolingual German runs

1.2.4.2 English Monolingual Retrieval

In the baseline experiment for English (EN-token) we find 617 relevant documents (out of 956; cf. table 2). The mean average precision (mAvP) is 0.35, and the average precision in the top ranks is high (AvP = 0.80). The difference between EN-token and EN-token-lemma is surprisingly small. This is due to the fact that English has fewer injected forms and hardly any noun compounding. Interestingly, using English lemmas decreases the precision. In section 1.2.4.4 we investigate the separate use of lemmas and tokens.

The performance level for English monolingual retrieval is significantly higher than for German. But when we add semantic indexing features, the general tendency in English monolingual retrieval is similar to German. MeSH leads to the best results both in recall and precision, UMLS is second best, and the semantic relations have almost no impact. The use of EuroWordNet terms (as it stands without word sense disambiguation) has a strong negative influence on the retrieval precision.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
EN-token	0.3455	617	0.8077	0.6160
EN-token-lemma	0.3320	635	0.7543	0.5760
EN-token-lemma-EWN	0.2565	616	0.6025	0.4640
EN-token-lemma-UMLS	0.3415	641	0.7516	0.5840
EN-token-lemma-MeSH	0.3543	648	0.7748	0.6000
EN-token-lemma-Semrel	0.3272	637	0.7279	0.5520

Table 2: Results of the monolingual English runs

1.2.4.3 Lemmas vs. Tokens for German

The previous experiments were based on the assumption that the combination of tokens and lemmas would naturally improve the retrieval quality. In a separate series of evaluation runs we checked the use of lemmas separate from tokens. Experiments based on linguistic stemming were carried out which we termed DE-lemma. In table 3 we present the results of these runs for the monolingual German retrieval experiments.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE-token	0.1600	322	0.5622	0.4160
DE-lemma	0.2809	591	0.6759	0.5320
DE-token-lemma*	0.2547	594	0.6744	0.5120
DE-lemma-EWN	0.2414	584	0.6140	0.4880
DE-lemma-UMLS	0.2754	590	0.6468	0.5200
DE-lemma-MeSH	0.2873	601	0.6647	0.5280
DE-lemma-Semrel	0.2795	591	0.6474	0.5200

Table 3: Results of the monolingual German runs

The baseline experiment for German (DE-token) is the same as in section 1.2.4.1.

The impact of linguistic stemming and decompounding is shown by the second experiment (DE-lemma), which achieves a recall gain of 70% compared to DE-token. In parallel, the precision figures have improved to an even higher level than for DE-token-lemma in table 1. But this time lemmatization was done in two steps. First we used the same general-purpose (i.e. general vocabulary) morphological analyzer as in the previous experiments. We observed that many medical terms were not lemmatized since they were not in the analyzer's lexicon. Therefore we developed heuristics for treating words that were unknown to the analyzer.

Based on these heuristics unknown adjectives were lemmatized by suffix truncation (e.g. *arthroskopischen* > *arthroskopisch*), and unknown nouns were decompounded if both compound parts were found as separate words in the corpus (*Nociceptinspiegel* > *Nociceptin Spiegel*). In this way the corpus itself was used as domain specific lexicon for decompounding. We can also use the lemma information of the second compound part if it is available in the corpus. For example, we can segment *Pertussisantigene* into *Pertussis* and *Antigene* since these two words occur stand-alone in the corpus. And we can then lemmatize the plural form *Antigene* into *Antigen* since this pair occurs in the corpus. These heuristics lead to 28,341 new adjective lemmas and 20'876 new noun lemmas from decompounding over all German documents.

We need to compare the result of DE-lemma with the combination of token and lemmas. Both were combined as indexing terms of equal weights in the queries and the documents. This combination leads to a decrease in precision (see DE-token-lemma*)

and therefore the tokens were discarded in the subsequent runs.

The impact of the different types of semantic information was determined one by one, but always in combination with lemmas. We wanted to support the hypothesis that semantic information will improve the precision over pure lemma information. The results show that the MeSH codes are the most useful indexing features. Using MeSH codes increases recall (from 591 to 601) and also average precision (from 0.2809 to 0.2873). As was to be expected the very specific semantic relations (Semrel) have hardly any impact. Using the EuroWordNet terms in combination with the lemmas degrades the overall performance.

1.2.4.4 Lemmas vs. Tokens for English

The baseline experiments for English (EN-token and EN-token-lemma) are the same as in section 1.2.4.2. But we can observe here that linguistic lemmatization (stemming) worsens the precision for English monolingual retrieval. It does increase the recall when used in combination with tokens (see line EN-token-lemma). This is very different from German monolingual retrieval, which clearly improves with lemmatization both for recall and precision.

The impact of the different types of semantic information was then determined in combination with tokens only. It turns out that again the MeSH codes are the most useful indexing features.

Using MeSH codes slightly increases recall (from 617 to 637) but mostly improves average precision (from 0.3455 to 0.3637). The impact of the UMLS terms is not visible and the very specific semantic relations (Semrel) have hardly any impact.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
EN-token	0.3455	617	0.8077	0.6160
EN-lemma	0.3097	600	0.6632	0.5360
EN-token-lemma	0.3320	635	0.7543	0.5760
EN-token-EWN	0.2155	604	0.5847	0.4000
EN-token-UMLS	0.3455	617	0.8077	0.6160
EN-token-MeSH	0.3637	637	0.8259	0.6040
EN-token-Semrel	0.3339	618	0.7555	0.5880

Table 4: Results of the monolingual English runs

Using the EuroWordNet terms in this combination with tokens degrades the overall performance. We investigated this phenomenon and found that EuroWordNet terms in our queries are mostly general language words like *injury, complication* or *treatment*. By using these words as additional indexing features we give them more weight than content-bearing specific terms. In a query like

Treatment of psychosomatical patients

we find the EuroWordNet terms *treatment* and *patient* while the most important word *psychosomatical* goes without notice. This leads to a bias towards the general language words and thus to a loss in retrieval precision.

1.2.4.5 Evaluation with the CMU relevance assessments

As mentioned in section 1.2.1 two sets of relevance assessments were produced by separate groups of medical experts. All retrieval results in the previous sections were based on the ZInfo relevance assessments. In order to check the influence of the relevance assessments on the retrieval results we repeated the monolingual German evaluation runs and checked them against the CMU relevance assessments. The results are listed in table 5.

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
DE-token	0.1410	193	0.4380	0.2320
DE-lemma	0.1664	283	0.3671	0.2600
DE-token-lemma	0.1785	291	0.4648	0.2840
DE-token-lemma-EWN	0.1658	287	0.4572	0.2880
DE-token-lemma-UMLS	0.1866	291	0.4629	0.3000
DE-token-lemma-MeSH	0.2001	296	0.4879	0.3120
DE-token-lemma-semrel	0.1767	291	0.4520	0.2840

Table 5: German monolingual retrieval with CMU relevance assessments

Since the total number of relevant documents in the CMU relevance assessments is only 500 as compared to 956 in the ZInfo assessments, the number of relevant retrieved documents must be lower but cannot directly be compared. And in parallel, the figures for P10 must be lower. If we have less relevant documents per query, then it gets more difficult to place them among the 10 top ranked documents.

But overall the tendency that we observed with the ZInfo relevance assessments is exactly reproduced with the CMU assessments: Lemmas do improve the retrieval results tremendously for German (cf. DE-lemma and DE-token-lemma). MeSH leads to the best results among the semantic codes. It improves recall slightly but precision on all measures.

1.2.5 Cross-Language Evaluation Runs

The cheapest way of Cross-Language Information Retrieval is monolingual retrieval over a parallel corpus. This means that we would search German documents with a German query and simply display those English documents that are known to be correspondences of the found German documents. This is not what we do here. Instead, we assume that we have a document collection (i.e. a corpus) in one language and a query in another language.

For most of the cross-language evaluation runs we used German queries to retrieve English documents. These results should not only be compared to the monolingual runs but also different approaches should be evaluated. All CLIR experiments were performed with the ZInfo relevance assessments.

1.2.5.1 CLIR via Vocabulary Overlap

A rough baseline for the cross-language task 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 in.ection (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.

1.2.5.2 CLIR via 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 to restrict the subject domain of the translation, and we selected the domains medicine and chemistry. This domain restriction helps the system to choose the subject-specific interpretation if multiple interpretations for a given lexical entry are available.

Although PersonalTranslator contains medical vocabulary, many words from our 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 for 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.

1.2.5.3 CLIR via Semantic Codes

Now let us compare these results with the semantic codes annotated in our corpus 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

1.2.5.4 CLIR via a Bilingual Similarity Thesaurus

For the further experiments Eurospider has built a similarity thesaurus (SimThes) over the parallel corpus [Qui 1995]. The similarity thesaurus contains words (adjectives, nouns, verbs) from our 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 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:

Similarity Thesaurus: *infarction, acute myocardial infarction, myocardial, thrombolytic, acute, thrombolysis, crs, synchronisation, cardiogenic shock, ptca*

	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-Xerox-SimThes (1)	0.3259	595	0.6910	0.6000
DE2EN-Xerox-SimThes (5)	0.3142	673	0.6763	0.5280
DE2EN-Xerox-SimThes (10)	0.2821	681	0.6064	0.4840
DE2EN-Xerox-SimThes (20)	0.2784	665	0.6049	0.4960

Table 9: CLIR results using a similarity thesaurus

We used these words for cross-language retrieval. Each German word from the queries was substituted by all the words of its similarity set. The similarity thesaurus is thus used for translation and query expansion. This resulted in a recall of 409 relevant documents found and a relatively good mean average precision of 0.2290 (see DE2EN-SimThes (10) in table 9).

Note that unlike in our previous experiments, we have now exploited the parallelism of the documents in our corpus for the construction of the similarity thesaurus. The bilingual similarity thesaurus is only available if we have a parallel or comparable corpus (cf. [Braschler and Schäuble 2000]) whereas the semantic annotations will also be applicable for a monolingual document collection.

We also checked the combination of all semantic annotations with the similarity thesaurus. Each query is now represented by its EuroWordNet, UMLS, MeSH and semantic relations codes as well as by the words from the similarity thesaurus. This combination leads to an even better precision for CLIR. We retrieved 518 relevant documents with a mean average precision of 0.2955 (cf. the line DE2EN-SimThes+all-

combined in table 9). And the figures for the high precision area (AvP and P10) are also very good. This means that a similarity thesaurus and semantic annotations complement each other.

For comparison we used another similarity thesaurus built by Xerox [Gaussier et al. 2000]. The main difference with respect to the Eurospider similarity thesaurus lies in the size of the context considered for retrieving translation equivalents: a pair of aligned sentences in the Xerox case, a pair of documents (or clusters of documents) in the Eurospider case. The lines DE2EN-Xerox-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 areused. Using between 5 and 10 similar documents thus seems like a good compromise between optimal precision and recall.

1.2.5.5 CLIR with English queries and German documents

In the above cross-language experiments we used German queries to find English documents. In addition we also evaluated the opposite translation direction. We used the English queries to obtain German documents. In an ideal setting the results should be exactly the same, since the documents are parallel (i.e. translations) and the queries are parallel. But of course even the best translation cannot guarantee a perfect transfer of the content.

Furthermore the MUCHMORE annotation tools and resources are language-specific. This results in a much denser semantic annotation for the English documents and English queries than for the German documents and German queries. For example, the English query *Associated diseases with insulin dependent diabetes mellitus* results in 8 semantic relations while the German equivalent results in only 3 semantic relations (2 of which overlap).

	mAvP	Rel. Docs Retr.	AvP 0.1	P10
EN2DE-EWN	0.0044	99	0.0197	0.0120
EN2DE-UMLS	0.1327	317	0.3892	0.2800
EN2DE-MeSH	0.1512	275	0.4372	0.3320
EN2DE-Semrel	0.0271	18	0.0800	0.0440
EN2DE-all-combined	0.1528	338	0.4279	0.3320
EN2DE-SimThes (10)	0.2259	492	0.5583	0.4840

Table 10: CLIR results of English to German

The results for the translation direction English to German are worse than for the opposite direction. But the relative usefulness of the various semantic codes are the same. MeSH scores highest in terms of precision but UMLS is better with respect to recall. Using a similarity thesaurus leads to the best results (cf. table 10).

1.2.6 Different weighting schemes

The *relevancy* information retrieval system allows for different methods of tuning the search process.

1. **Coordination level matching**. Coordination level matching means that the weight of the documents is not calculated with a classical relevancy scheme. Instead we consider the neighbourhood of the query terms in the document and their position in the document as primary relevance criterion. The smaller the window in the document containing all query terms and the nearer this smallest window is from the document start so much the better its relevance is weighted.

2. **High priority and low priority terms**. When using multiple indexing terms they can be split into high and low priority terms.

3. **Weighting factors for different indexing features**. The weighting method tells the system how to calculate the relevance values for the documents.

All of these weighting schemes were explored alone and in combination. But due to the short queries and the small document collection (less than 10,000 documents) these schemes did not result in any improvement of the retrieval results. Therefore they were not used in any of the above-mentioned experiments.

2 CMU Evaluation

In this section of the report we present first, in subsection 2.1, the concept-based classification approach that uses the MeSH hierarchy. Then the corpus-based approaches used are presented in subsection 2.2. Results of the retrieval experiments with these approaches are presented in subsection 2.3.

2.1 Hierarchical MeSH Concept Classification

2.1.1 Overview

As noted earlier, the essential part of any concept-based CLIR system is the mapping of terms to a language-independent conceptual level. Our classification-based approach relies on the documents in the search space being labeled in accordance with the language-independent 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.

2.1.2 Retrieval With MeSH Classification

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.

2.2 Corpus-Based Approaches

For comparison purposes, we also worked with two corpus-based approaches, Example-Based Thesaurus and Pseudo-relevance Feedback, as described in the following two subsections.

2.2.1 Example-Based Thesaurus

The Example-based Thesaurus (EBT) approach 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].

2.2.2 Pseudo-Relevance Feedback

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, 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 $TF * IDF$.

2.3 Evaluation

The evaluation numbers presented in Table 11 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 labeled 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, labeling the German half used the German half of the Springer training set, thus suffering two levels of machine assignment.

However, as shown in Table 11, 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 11: 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.)

3 Conclusions

We have explored the use of different kinds of semantic annotation for both monolingual and cross-language retrieval. We have also explored concept-classification based retrieval and compared it with traditional term-based retrieval.

In monolingual retrieval (for both English and German) semantic information from the MeSH codes (Medical Subject Headings) were most reliable and resulted in an increase in recall and precision over token and lemma indexing. Moreover, the monolingual experiments show that high-quality linguistic analysis is crucial for a good retrieval performance, which indicates that further work is needed to improve the compatibility and quality of morphological analysis both on the side of document and query processing and indexing. This is a prerequisite for a good baseline.

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.

Second, semantic codes from UMLS and MeSH can be safely combined for CLIR. By using these codes we can reach a level comparable to Machine Translation without the domain-specific lexicon.

Semantic codes were 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.

This result was comparable to the German monolingual retrieval results. Using a similarity thesaurus built over on sentence alignment fared even better and led to the best results. If we compare the monolingual and cross-language retrieval results, it is striking that the best semantic sources in the monolingual experiments were also the best in the cross-language task. This indicates that monolingual results for semantic annotations can

be extrapolated to cross-language retrieval if no cross-language test set is available.

For MeSH-classification-based retrieval, the combination of term-based and concept-based approaches produced an improvement in all cases except English monolingual.

In summary, best results in CLIR were consistently obtained by use of both corpus-based and concept-based approaches in combination.

References

[Ballesteros and Croft 1997]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.

[Braschler and Schäuble 2000] Martin Braschler and Peter Schäuble. 2000. Using corpusbased approaches in a system for multilingual information retrieval. *Information Retrieval*, (3):273–284.

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

[Brown 1997] 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.

[Buitelaar and Sacaleanu 2001] P. Buitelaar and B. Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proc. Of NAACL WordNet Workshop*, Pittsburgh.

[Carbonell et al 1997] 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.

[Eichmann et al. 1998] D. Eichmann, M. Ruiz, and P. Srinivasan. 1998. Cross-language information retrieval with the UMLS metathesaurus. In *Proc. Of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia.

[Gaussier et al. 1998] E. Gaussier, G. Grefenstette, D.A. Hull, and B.M. Schulze. 1998. Xerox TREC-6 site report: Cross language text retrieval. In *Proc. Of the Sixth TExt Retrieval Conference (TREC-6)*, Gaithersburg, MD. National Institute of Standards Technology (NIST).

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

[Hersh et al. 1994] 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.

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

[Raileanu et al. 2002] D. Raileanu, P. Buitelaar, J. Bay, and S. Vintar. 2002. Evaluation corpora for sense disambiguation in the medical domain. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las

Palmas, Canary Islands, Spain, May 29-31.

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

[Schäuble and Sheridan 1998] Peter Schäuble and Patrick Sheridan. 1998. Cross-language information retrieval (CLIR). track overview. In *Proc. Of The Sixth Text Retrieval Conference (TREC-6)*, Gaithersburg, MD. National Institute of Standards Technology (NIST).

[Schäuble 1997] Peter Schäuble. 1997. *Multimedia Information Retrieval. Content-based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, Boston.

[Srinivasan 1996] Padmini Srinivasan, "Optimal Document-Indexing Vocabulary for MEDLINE", Information Processing & Management, 32(5), pp. 503-514, 1996.

[Vintar et al. 2002] S. Vintar, P. Buitelaar, B. Ripplinger, B. Sacaleanu, D. Raileanu, and D. Prescher. 2002. 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.

[Volk and Buitelaar 2002] Martin Volk and Paul Buitelaar. 2002. A systematic evaluation of concept-based cross-lingual information retrieval in the medical domain. In *Proc. Of 3rd Dutch-Belgian Information Retrieval Workshop*, Leuven.

[Volk et al. 2002] M. Volk, B. Ripplinger, S. Vintar, P. Buitelaar, D. Raileanu, and B. Sacaleanu. 2002. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67(1-3):97–112, December.

[Wechsler et al. 1997] M. Wechsler, P. Sheridan, and P. Schäuble. 1997. Multi-language text indexing for internet retrieval. In *Proc. Of the RIAO'97 Computer-Assisted Information Searching on the Internet*, Montreal, Canada.

[Yang et al. 1998] 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.

The Set of Queries

	English	**German**
1.	Arthroscopic treatment of cruciate ligament injuries	Arthroskopische Behandlung bei Kreuzbandverletzungen
2.	Complications of arthroscopic interventions.	Komplikationen bei arthroskopischen Eingriffen
3.	Pathophysiology and prophylaxis of arthrofibrosis	Pathophysiologie und Prävention der Arthrofibrose
4.	HIV Epidemiology, Risk Assessment	HIV Epidemiologie, Risikoabschätzung
5.	Patient-controlled analgesia indications and limits	Patientengesteuerte Analgesie, Indikationen und Grenzen
6.	Priming with non-depolarizing muscle relaxants	Priming mit nicht-depolarisierenden Muskelrelaxanzien
7.	Complications after laparoscopic cholecystectomy	Komplikationen nach der laparoskopischen Cholecystektomie
8.	Heparin induced thrombocytopenia, diagnosis and management	Heparininduzierte Thrombozytopenie, Diagnostik und Handhabung
9.	Diagnostic in Lyme disease.	Diagnostik der Lyme-Borreliose.
10.	Treatment of acute myocardial infarction.	Behandlung des akuten Myokardinfarkts.
11.	Catheter ablation and cardiac mapping.	Katheterablation und kardiales Mapping.
12.	Diagnostic approach in injuries of the shoulder.	Diagnostische Ansätze bei Verletzungen der Schulter.
13.	Differential diagnosis in infertility.	Differentialdiagnostik bei Unfruchtbarkeit.
14.	Approach of the correction of deformities in orthopedics.	Möglichkeiten der Korrektur von Deformitäten in der Orthopädie.
15.	Treatment of squamous cell carcinoma	Behandlung von Plattenepithelkarzinomen.
16.	Associated diseases with insulin dependent diabetes mellitus	Begleitende Erkrankungen von Insulin abhängigen Diabetes mellitus.
17.	Therapy in chronic low back pain	Therapie bei chronischem Rückenschmerz.
18.	Treatment of ventricular tachycardia	Behandlung der ventrikulären Tachykardie.
19.	Indication for implantable cardioverter defibrillator (ICD)	Indikation für einen implantierbaren Kardioverter-Defibrillator (ICD).
20.	Diagnostic in acute and chronic myocarditis	Diagnostik der akuten und chronischen Myokarditis.
21.	Cause of dysphagia	Ursachen von Schluckstörungen.

22.	Treatment of sensorineural hearing loss (SNHL)	Behandlung des sensoneuralen Hörverlust (SNHL).
23.	Complications of surgical repair of aortic aneurysm	Komplikationen bei der chirurgischen Therapie von Aortenaneurysmen.
24.	Treatment of psychosomatical patients	Behandlung von psychosomatischen Patienten.
25.	New approach in cruciate ligament surgery	Neue Erkenntnisse in der Kreuzbandchirurgie.

Comparison of the relevance assessments

	Query	*ZInfo*	*CMU*	*both*	*only ZInfo*	*only CMU*
1	Arthroscopic treatment of ruciate ligament injuries	51	16	13	38	3
2	Complications of arthroscopic interventions.	17	29	13	4	16
3	Pathophysiology and prophylaxis of arthrofibrosis	24	4	3	21	1
4	HIV Epidemiology, Risk Assessment	39	6	6	33	
5	Patient-controlled analgesia indications and limits	44	5	4	40	1
6	Priming with non-depolarizing muscle relaxants	37	4	4	33	
7	Complications after laparoscopic cholecystectomy	20	10	10	10	
8	Heparin induced thrombocytopenia, diagnosis and management	19	17	16	3	1
9	Diagnostic in Lyme disease.	15	9	8	7	1
10	Treatment of acute myocardial infarction.	66	35	29	37	6
11	Catheter ablation and cardiac mapping.	42	21	20	22	1
12	Diagnostic approach in injuries of the shoulder.	18	21	12	6	9
13	Differential diagnosis in infertility.	25	27	17	8	10
14	Approach of the correction of deformities in orthopedics.	75	74	55	20	19
15	Treatment of squamous cell carcinoma	39	21	14	25	7
16	Associated diseases with insulin dependent diabetes mellitus	29	16	12	17	4
17	Therapy in chronic low back pain	35	33	23	12	10
18	Treatment of ventricular tachycardia	104	26	26	78	
19	Indication for implantable cardioverter defibrillator (ICD)	25	26	15	10	11

20	Diagnostic in acute and chronic myocarditis	7	8	4	3	4
21	Cause of dysphagia	22	19	13	9	6
22	Treatment of sensorineural hearing loss (SNHL)	67	18	16	51	2
23	Complications of surgical repair of aortic aneurysm	32	10	9	23	1
24	Treatment of psychosomatical patients	49	19	15	34	4
25	New approach in cruciate ligament surgery	58	26	25	33	1
		956	500	382	577	118