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 06 = December, 2000
Actual date of delivery	January, 2001
Deliverable number	D1.1
Deliverable name	State of the Art in Cross-lingual Information Access for Medical Information
Type	*Report*
Status & version	*Final*
Number of pages	130
WP contributing to the deliverable	WP 1
WP / Task responsible	CSLI
Other contributors	DFKI, XRCE, Zinfo, EIT, CMU, CSLI
Author(s)	(see table below)
EC Project Officer	Yves Paternoster
Keywords	Term Extraction; Relation Extraction; Text Summarization; Cross-language Information Retrieval; Word Sense Disambiguation; Medical Information
Abstract (for dissemination)	This report provides an overview of available and emerging techniques and resources and their theoretical underpinnings in the areas of concept-based information access, cross-language information retrieval, word sense disambiguation, and medical domain modelling. Special attention is paid to the use and availability of linguistic data for supervised and unsupervised methods, benchmarks and standards of evaluation, and the nature and specific challenges of medical information flow, terminology and ontology.

Contact Information:

MUCHMORE : DFKI GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany

Prof. Dr. Hans Uszkoreit, General Manager
Dr. Paul Buitelaar, Technical Coordinator
Dr. Michelle Carnell, Financial Coordinator

Editors: CSLI, Ventura Hall
220 Panama Street
Stanford, CA 94305-4115

Stanley Peters, Professor of Linguistics
Stefan Kaufmann, Graduate Student of Linguistics

Contributors:

SECTION	AUTHOR(S)	
1.1 Terminology Acquisition	David Hull	XRCE
1.2 Relation Extraction	Paul Buitelaar	DFKI
1.3 Text Summarization	Jaime Carbonell, Jade Goldstein	CMU
2.1 Dictionary-based, Corpus-based and Concept-based Approaches	Yiming Yang	CMU
2.2 State of the Art in Parallel-corpus Based Methods	Ralf Brown	CMU
2.3 Non-parallel (Comparable) Corpora	Martin Braschler	Eurospider
2.4 Evaluation	Martin Braschler	Eurospider
3 Word Sense Disambiguation	Paul Buitelaar, Stefan Kaufmann	DFKI, CSLI
4 The Medical Domain	Jörg Bay, Christoph Winkler, Oktavian Weiser	ZInfo

Table of Contents

List of Tables

List of Figures

1 Summary

The focus of the MUCHMORE project is on cross-lingual information access, applied to the domain of medical information management. Our work will draw on and contribute to a number of different research areas in linguistics and natural-language processing, each with its own specific challenges and corpus of prior work.

This report provides a survey of those areas, existing results, techniques and evaluation standards, with special attention paid to their relevance to the project and its intended domain of application. It serves both to ensure that the project draws maximal benefit from previous work and available resources, thus avoiding duplication of work, and as a medium for the participants to share their expertise in their respective areas of strength in order to arrive at a common understanding of the nature and complexity of the tasks involved. Furthermore, by making the report publicly accessible, we hope that it may serve as a resource for other researchers interested in the area.

We identified four major topics and structured the document accordingly. There is, however, some overlap between the problems arising in each of these areas, and a number of themes - such as the distinction between supervised and unsupervised methods or that between parallel and comparable corpora, and the measures of precision and recall in evaluation - recur in different parts of the report.

The first chapter on concept-based information access deals with the identification of meaningful items and patterns in free, unannotated text. The material on which our application is intended to operate will be medical documents in a variety of genres, styles and languages. The automatic detection and extraction of terminology and semantic relations in such texts is an essential step toward the ability to use existing domain models in their categorization and processing. The section on text summarization deals with ways of presenting brief overviews of the contents of long documents in a user-friendly and useful way.

The second chapter is devoted to the related area of cross-language information retrieval. An important goal of the MUCHMORE project is the ability to search large, multilingual document collections in response to concise, monolingual user queries. In this context, the need to "transfer" information between languages adds to the complexity of the information retrieval. The chapter discusses approaches to this task that rely on a variety of resources, as well as ways of producing those resources.

Word sense disambiguation, the topic of the third chapter, addresses the fact that most words in natural languages can be used with more or less radically distinct meanings in different contexts. This flexibility usually goes unnoticed in human use, but is a major source of errors and complexity in computational applications. Determining which sense a given occurrence of a word has is an important enabling task for the higher-level semantic processing discussed in the earlier chapters.

Finally, the fourth chapter provides a survey of the specific problems involved in organizing the domain of medical knowledge and its terminology. It includes discussions of the ways in

which information is structured and used in the medical professions, the various attempts at codifying and standardizing medical terminology underway in large-scale projects, as well as the available resources, such as thesauri, nomenclatures and dictionaries, produced in those efforts.

2 Concept-based Information Access

2.1 Multilingual Terminology Acquisition

2.1.1 Introduction

A term is usually described as a lexical unit (word or phrase) that has a stable (usually specialized) meaning in a particular domain. Terminology dictionaries are an important resource for controlled authoring, translation, indexing for text retrieval, and many other language processing tasks (knowledge representation, expert systems, etc.). Terminology databases are traditionally built by hand, a time-consuming, resource-intensive, and often tedious task which must be performed by experts in the field. In response to this challenge, there has been a lot of work in the last few decades on automatic (or semi-automatic) terminology extraction from corpora. More recently, this work has been extended to multilingual texts, where systems also search for term translations.

In this section, we review the existing research and technology in the area of terminology acquisition and relate it to the needs of the MUCHMORE project. Rather than conduct an exhaustive survey of this area, we will provide a more general overview and then focus attention on the developments particularly relevant to the project. The section is divided into two main parts: monolingual and multilingual terminology acquisition, reflecting a natural separation of the problem space. Monolingual acquisition addresses the problem of automatic terminology extraction, while multilingual acquisition is more concerned with finding translations. Most terminology extraction systems rely on some linguistic components, which means that there are many language dependent issues. We will focus our attention on German and English, the primary langauges of interest for the MUCHMORE project, and touch on issues relating to the medical domain. The subject of medical terminology will be covered in more depth in Chapter 4. The key issue in multilingual documents is the degree of alignment between the documents in the different languages (i.e. simultaneous transcription, direct translation, indirect translation, degree of domain overlap, etc.) This section is divided into two parts, covering parallel and comparable corpora.

It is important to recognize from the start that there are strong upper bounds on the performance level of automatic terminology acquisition systems. This is due to the fact that the definition of a terminology unit and the boundaries of the subject field to which it is relevant are often unclear. They will vary from one domain expert to another and from one application to another. Since terminology extraction can never be a fully automatic process, input from a human expert will be required to attain very high accuracy.

The easiest way to come up with a clear notion of a term is to compare the usage of technical terminology to that of general language words. A technical term has a fixed meaning,

regardless of context. In contrast, the meaning of a word in a general language setting is highly dependent on its context. A term can be taken out of context and its meaning will remain clear and unequivocal. Terms which represent different concepts will sometimes share the same surface form, but this is very different from the subtle gradations of meaning associations with general language words. Multi-word terms often have a meaning which is non-compositional, i.e. the expression has a very different sense than that of the individual words taken together. General language expressions, other than idioms and certain lexical units that happen to written as two words, very rarely have this property. Finally, terms have a clear place in an ontology or semantic network covering the domain.

In order to understand the key role of terminology in the MUCHMORE project, we must show why terms are important for information retrieval. In the recent past, IR systems worked exclusively with a set of controlled indexing units. In this case, there is a perfect mapping between terminology and indexing units. Both describe the important concepts of the domain, as identified by human experts. However, modern IR systems have moved towards a much more exhaustive indexing of document collections. The accepted procedure is to index the full text of all documents, except for a small set of close-class function words. Many IR systems index multi-word units as well or can test for the presence of such units using boolean proximity operators. This movement reflects two basic trends. First, IR systems have been opened to the masses by the spread of electronic information, particularly via the World Wide Web. Searchers who are not domain experts cannot be expected to use a controlled list of indexing terms effectively. Second, computational power and storage space have increased dramatically, making full-text indexing and retrieval possible. With full-text indexing, the clear link between terminology and indexing units has disappeared.

However, terminology recognition is still an important issue for information retrieval in specialized domains, as terminology can add important meta-information to the raw document content. Documents will often be assigned indexing terms that never occur in the text itself. Automatic full-text indexing systems do not directly address synonymy or other important semantic relations. Controlled indexing terms have a consistent meaning, and are automatically normalized, addressing two key problems in automatic indexing, polysemy and syntactic variation. Thesauri and semantic networks can be a particularly valuable resource even when a document collection has not been manually indexed, for they enable query expansion for full text retrieval. Automatic query expansion techniques tend to produce a lot more noise. For these reasons, terminology databases serve as valuable supplemental resources for information retrieval.

The justification for fully automatic terminology extraction without manual verification for text indexing is less clear. IR experiments have not shown conclusively that such an approach is better than simply indexing adjacent word pairs. On the other hand, most forms syntactic variation can be normalized automatically, and a lot of it extends beyond a window of length two. However, terms are much more relevant for cross-language text retrieval, because they are the core semantic units for the domain, and thus the most appropriate unit of translation. This is particularly important for multi-word compounds with non-compositional meaning. Unless they are clearly recognized as terms, there is no way to translate them correctly.

2.1.2 Automatic Terminology Extraction

The process of automatic terminology extraction can be divided into the following basic steps:

1. Candidate term extraction:
   identify the morpho-syntactic patterns associated with potential terms

2a. Term filtering:
   sort candidates according to the likelihood that they are terms (separate terms from idioms, proper names, and other fixed lexical units)

2b. Sub-term extraction:
   for any given multi-word term, extract all sub-sequences that are potential terms

3. Term clustering and networking:
   recognize morpho-syntactic variants and group them together, identify links between terms (most common types: synonymy, hyperonomy, hyponomy)

2.1.2.1 Candidate Term Extraction

The strong relation between morpho-syntactic patterns and terminology was first exploited in depth for automatic terminology extraction by the TERMINO system at the Université de Québec à Montréal (David and Plante, 1990; Lauriston, 1994). It is commonly assumed that the vast majority of technical terms are noun phrases (NP's). For example, Justeson and Katz (1995) reports that 94 of 97 unique term types (as opposed to tokens) in a particular technical article are NPs. Similarly, Arppe (1995; Table 2) shows that at least 95% of the syntactic patterns for terms (in a sample of 558 types) are NPs, and it is likely that most of the remaining 5% are as well. Therefore, 95-99% recall can be attained for many corpora simply by extracting all noun phrase patterns. However, these numbers are highly corpus/domain/language-dependent, so they should be used with caution. For example, Heid *et al.* (1996) extracts many verbs and adjectives also classified as terms in a German automotive corpus.

Unfortunately, it is hard to build a highly accurate terminology extraction system on the basis of syntactic pattern matching alone, since many noun phrases are not terms. Arppe finds that NP extraction obtains an average precision of 27% (counting types), with the precision by syntactic pattern ranging from roughly 15-40% (the patterns NN and ANN have the highest precision [A = adjective, N = noun]). Justeson (1995) has slightly more success, with precision values of 67%, 73%, and 92% on three different articles, but much of this difference is likely due to the texts that were chosen. Most systems address this issue by trading recall for precision. Noise reduction strategies will be discussed in the next section on term filtering.

Another important issue is term decomposition. Most systems start by extracting the longest matching syntactic pattern and then parse it in order to identify potentially valid sub-terms. If all possible sub-terms are extracted, this can multiply the total number of term candidates by a significant factor. Therefore, good term filtering and selection algorithms are critical in this context. Term decomposition is necessary for information retrieval, because the likelihood of finding an exact match between a term in the query and the document decreases with the length of the term. Matching on term constituents is important for the robustness of the system.

While any realistic approach to terminology recognition must involve some kind of morpho-syntactic pattern matching, many people have worked on extracting word collocations using statistical methods for information retrieval and lexicography. Word collocation patterns have at least as much noise as morpho-syntactic patterns, but can serve as a valuable additional source of evidence to decide whether a particular word sequence is a likely term.

So far, we have dealt with terminology extraction in a language neutral fashion. Obviously, morpho-syntactic pattern matching requires language-specific lexical resources. In addition, languages have different approaches to term construction, depending on their origin. German is a compounding languages, meaning the multi-word terms are built by the direct concatenation of individual words (i.e. without prepositions). German compounds are single-word units, making term boundary recognition simple. English is primarily a compounding language, although the units of a compound are separated by spaces, making term boundary recognition somewhat more difficult. Some terms contain the preposition of, but this a relatively infrequent phenomenon. However in English, a large fraction of nouns can be used as verbs and share the same surface form. This makes part of speech disambiguation for terminology recognition much more challenging. In contrast to German and English, Romance languages, such as French, use prepositions for term composition. This creates a lot of ambiguity between multi-word terms and prepositional phrases, making the term recognition problem much more difficult. Fortunately, the MUCHMORE project focuses primarily on German and English.

2.1.2.2 Term filtering

The goal in term filtering is to reduce the noise associated with term extraction by syntactic patterns. Strategies for term filtering can be broadly divided into two categories: statistical and linguistic. Linguistic methods involve further refinements of the syntactic patterns used in the term candidate extraction step, often drawing on lexical information. Statistical methods generally compute a score measuring the degree of “termhood.” Candidate terms are then ranked by their score, allowing the most likely terms to be analyzed first. Alternatively, the list can be truncated, trading recall for precision.

Broadly speaking, we distinguish between two types of linguistic filters, lexical filters and syntactic filters. Lexical filters consist of a list of words not commonly associated with terms. Some of these words may be associated with a particular domain while others cut across all domains. Syntactic filters consist of exception rules which identify syntactic patterns rarely associated with terms. Exception rules will sometimes eliminate valid terms, but this is often justified in order to improve the precision of the candidate set of terms. Exception rules differ from the syntactic patterns described in the previous section in that they are built on top of existing extraction rules in order to improve the accuracy of the system.

A variety of statistical scoring functions have been proposed. The most basic approach is to measure raw frequency. The more often a candidate appears in the corpus, the more likely it is to be a term. A slightly more sophisticated approach is Ahmad's weirdness coefficient (Ahmad, 1994,) which is simply the ratio between the frequency of a word in a given domain and its frequency in general language. This scoring strategy requires a large collection of documents considered representative of general language. For terms of length two, there are wide range of statistical association measures which compare the observed frequency of collocation with the value expected by chance. Daille (1994) explores eighteen different association measures, finding that the binomial likelihood ratio test is the most appropriate for selecting terms. Focusing attention on terms of length two may seem rather restrictive at first,

but the vast majority of longer terms can be derived from terms of length two. Binary association measures can be used on longer terms in a recursive manner.

2.1.2.3 Sub-term Extraction

An important related issue in term filtering is the extraction of sub-terms (i.e. sub-sequences of longer terms). While one can follow the same pattern matching strategy described above, this can lead to massive over-generation. For example, a term of length *n* can have as many as 2^n possible sub-terms, which can be quite a problem for a term of length 9 such as:

```
(1)  light weight unitary copolymiric plastic removable rear
     canopy section
```

extracted from a patent document. If attention is restricted to adjacent sub-sequences, this still yields $n*(n+1)/2$ possible sub-terms. Therefore, complete sub-term extraction is rarely feasible without some kind of selection method. The simplest selection method is to retain only sub-terms which are represented as independent terms elsewhere in the collection. This method removes a great deal of noise, but may also eliminate many valid terms. A slightly more sophisticated approach is to measure productivity. Productivity is usually defined as the number of different words either preceding or following a given candidate term. Candidates with high productivity are more likely to be useful terms. Frantzi and Aniandou (1997) capture this notion with a measure known as "C-value":

Eq 1: C-value

$$\text{C - value}(T) = (length(T) - 1)\left(freq(T) - \frac{t(T)}{c(T)} \right)$$

length(T)	= number of words in *T*,
freq(T)	= total frequency of *T* in the corpus,
t(T)	= frequency of *T* in longer candidate terms,
c(T)	= number of longer candidate terms including *T*

C-value is basically raw frequency with a penalty factor inversely proportional to the productivity of a term. In the extreme case, a sub-term which only appears in a single longer term will receive a score of zero. Another productivity measure is proposed by Nakagawa and Mori (1998.) Bourigoult *et al.* (1996) also use productivity to help determine the boundary of certain candidate terms.

Alternatively, one can attack sub-term extraction in a more general way by treating it as a noun-phrase parsing problem. In this setting, the goal is to recognize and bracket the strongest associations, e.g.

```
(2)  [[light weight] unitary copolymiric plastic removable
     [[rear canopy] section]]
```

This leads to a natural selection of sub-terms, i.e. `rear canopy` and `rear canopy section`, but not `rear section`. Association can be measured in exactly the same way as it is for term filtering. The key difference in this context is that it is the relative strength of association within the compound that is important. Zhai *et al.* (1997) uses this approach for extracting indexing terms for the CLARIT system. Lauer (1995) has evaluated a number of approaches to selection for the special case of NNN compounds. Zhai (1997) suggests a more sophisticated probabilistic model of noun phrase structure that maximizes the likelihood over all observed NPs simultaneously.

2.1.2.4 Term clustering and networking

Term clustering describes the process of recognizing term variants and grouping them together. Jaquemin (1996) recognizes three major types of variation: morphological, syntactic, and semantic. Morphological variation includes: inflection (`canopy` vs. `canopies`), agglutination/punctuation/hyphenation, (`co-polymiric` vs. `copolymiric`, `x ray` vs. `x-ray`), and abbreviation/acronyms. Jacquemin further distinguishes three types of syntactic variation: coordination (`front and rear canopy section`), insertion/elision (`neutral position` vs. `neutral balanced position`), and permutation (`nylon woven cloth` vs. `cloth of woven nylon.`) Semantic variation is a general category which includes synonyms that don't share syntactic structure.

Partial or shallow parsing techniques are effective for detecting syntactic variation, particularly for coordination and insertion. Jacquemin's FASTR system achieves 97-98% precision and 98% recall for terms with three content words for coordination and insertion. However, these figures drop to 68% precision and 95% recall for permutation. Term variation is a non-negligible phenomenon. Jacquemin reports that 15% of all multi-word terms are variants. Term variant clustering can be used for query expansion to increase the recall of the search results.

Term networking describes the process of linking together terms according to semantic relations. The most common types of semantic structure are synonymy, hyperonomy, and hyponomy. Automatically detecting semantic variation is a much more difficult problem, as many semantically related terms have little or no common syntactic structure. Therefore, semantic networks are always always built by hand. A good semantic network or hierarchy can be an extremely valuable resource for domain-specific information retrieval. However, some types of relations can be captured automatically using syntactic structure. The most common approach is to link together terms sharing common lexical units (head nouns, modifiers), but this is the least useful for information retrieval. More sophisticated terminology networks are usually built by hand, but there has been some interesting recent work on automatic and semi-automatic construction of semantic links.

Hearst (1992) identifies a set of syntactic expressions that are commonly associated with hyponyms and uses them to extract term pairs from corpora. For example, the phrase `. . . works by such authors as Shakespeare . . .` can be used to extract the relation: `hyponym("author", "Shakespeare")`. Morin (1999) bootstraps a list of potential patterns automatically from an existing semantic resource. For example, if

Shakespeare is known in advance to be an author, then extracting all sentences containing both `Shakespeare` and `author` provides a set of potential patterns which could help to identify other hyponyms. Morin extracts sentences from many known hyponyms and clusters them to identify common patterns. These patterns are then manually validated (to increase precision) before they are used for extraction. This method extracts 836 hyponyms with 79% precision and 40% recall on a technical corpus. Nédellec's (1999) Asium system takes verb sub-categorization data extracted by a robust parser and uses clustering algorithms combined with a specialized user interface to enable domain experts to rapidly construct a high-quality ontology.

The problem of automatic thesaurus construction has been heavily researched in information retrieval and computational linguistics. These methods are largely based on building a profile for each term and clustering terms on the basis of a profile similarity functions. Profiles can be generated from word co-occurrence information (Qiu and Frei, 1993; Schütze, 1997; Collier *et al.*, 1998; Xu and Croft, 2000,) or syntactic relations such as subject-verb-object extracted with a robust parser (Grefenstette, 1994; Lin, 1998.) Automatically-generated clusters contain a mixture of syntactic and semantic relations as well as a lot of noise (spurious co-occurrences). While such resources are generally not suitable for sophisticated language processing tasks, they have proven to be quite effective for monolingual and cross-language text retrieval. Thesauri are commonly used in information retrieval as tools for query expansion. In other words, all terms in the same concept (or co-occurrence) class as the query terms are added directly to the query. For CLIR, the query can be expanded across languages, serving as a form of translation.

Alternatively, term clustering and networking can be used to enhance existing semantic resources. Since the MUCHMORE project addresses the medical domain, where a substantial set of resources already exists (e.g. UMLS), this is the logical approach. In this model, the goal is to find variants of existing terms and classify new terms in the appropriate category in a thesaurus. This task is much more tractable than inducing structure completely from scratch. Given enough training data, it may be possible to automatically learn the syntactic patterns that induce semantic relations, following a model similar to the one suggested by Morin above. This will be an important area of research in the MUCHMORE project.

2.1.3 Multilingual Terminology Acquisition

The availability of multilingual corpora opens up new possibilities for terminology acquisition. Depending on the structure of the corpora, it may be possible not only to recognize terms, but also to search for their translations. We distinguish between two types of corpora for multilingual terminology acquisition (MTA): *parallel* corpora and *comparable* corpora. In computational linguistics, a parallel corpus is defined as a collection of documents and their translations. A comparable corpus is a collection of documents in several language from the same domain. Comparable corpora can also sometimes be aligned at the document level. A comparable corpus is aligned if it is possible to link two or more documents to each other on the basis of external data, such as date of publication, subject tags, or other meta-information. Linked documents will be closely related but not direct translations of each other. For example, two news articles written about the same event by different people in different languages would be considered aligned. There is good reason to expect that such articles will use a lot of common terminology and should be a rich source of terminology translations.

The goal in multilingual terminology acquisition is to recognize and extract terms and their translations; in other words, to build a bilingual or multilingual terminology lexicon. If no attempt is made to find translations, then the problem reduces to monolingual extraction as described in the previous section. Therefore, the focus of this section will be on alignment techniques rather than extraction techniques, although for some models, these two steps cannot be completely separated. The quality of the extracted lexicon depends strongly on the degree of alignment between the texts. With parallel corpora, systems can achieve a relatively high level of accuracy, on the order of 80-90% precision and recall. However, parallel texts are hard to find and usually represent only a small fraction of the available resources in any setting. Therefore, we will be forced to rely to a large extent on comparable corpus alignment techniques, which have much lower levels of accuracy. If these results are to be integrated into a cross-language text retrieval system, manual or semi-automatic validation may be required.

2.1.3.1 Parallel corpora

The task of terminology acquisition from parallel corpora will be called *bilingual terminology alignment*. In this section, we discuss the alignment of corpora which are parallel in two languages. These techniques can be applied to parallel corpora with more languages in a pairwise fashion. Terminology alignment of three or more languages simultaneously is a much more challenging problem, since many terms do not have exact equivalents in all languages. The algorithms described here generally work with a corpus that has already been aligned at the sentence level. This is not an unreasonable assumption, since many parallel corpora can be accessed through a translation memory (defined as a set of pairs of text segments, usually sentences, that are translations). If not, recent advances in automatic sentence alignment technology enable systems to build such resources quickly and accurately.

There are three major approaches to bilingual terminology alignment:

1. Independent extraction of terms in the source and target language, followed by alignment

   The advantage of this approach is that it is highly modular. The extraction and alignment algorithms function independently, so the system can be easily extended to new language pairs if the alignment algorithm is language independent. However, it suffers from resolution mismatch problems, due to differences in the extraction algorithms. For example, a concept may be represented by one term in language A and two terms in language B. If the goal is to find the translation for one of the terms in language B, the system is likely to fail unless good term decomposition algorithms for language A are also available. Kupiec (1993) and van de Eijk (1993) follow this approach.

2. Extraction in the source language, followed by alignment to the most appropriate text string in the target language

   This option may be the only feasible approach if terminology extraction algorithms are available only for the source language. It is also a good choice if the extraction accuracy is much higher in one language than another. For example, it is often used for English-French parallel corpora, since terminology extraction in English is a much simpler theoretical problem. Finally, this method has higher recall than method (1),

because it can extract a target language translation that is not itself a term. Smadja (1996), Hull (1997), and Gaussier (1998) have developed algorithms based on this approach.

3. Simultaneous extraction and alignment

   In some sense, this model is the most powerful, for it is capable of improving the accuracy of extraction in one language using information from the other language. However, current research has not fully exploited the potential of this model. One example of this approach is the inversion transduction grammar described in Wu (1995) which is designed explicitly for bilingual language modelling. While this method is very promising, the bilingual grammar is language independent, so it cannot exploit specific properties of the individual language. However, any language-dependent bilingual modelling scheme would be much more complex, and thus correspondingly difficult to port to new language pairs.

This section has focused on systems designed explicitly to solve the bilingual terminology alignment task. In addition, there has been a lot of important work on the more general task of machine translation that is relevant to this problem. Parallel corpora have become an important resource for statistical (Al-Onaizan *et al.*, 1999) and example-based (Brown, 1997) machine translation systems. While these techniques are not designed explicitly for terminology alignment, they could easily be adapted to the problem.

2.1.3.2 Comparable Corpora

In the introduction, we distinguished between two types of comparable corpora, aligned and unaligned. Aligned comparable corpora have been used in many of the recent TREC cross-language retrieval experiments, so a number of CLIR techniques have been developed based on this research. Among the techniques used are probabilistic translation models of the type used for statistical machine translation, latent semantic indexing, similarity thesauri, example-based machine translation, and the generalized vector space model. Most of these techniques will be described in the section of the report on the state of the art in cross-language text retrieval, so we will not present them again here.

Researchers have recently begun to build alignment algorithms based on unaligned comparable corpora. At present, we are aware of only three efforts in this area, Rapp (1995), Peters (1995), and Fung and Lee (1998). All of these systems share a common basic strategy. First, build term profiles for the source and target language. A profile consists of a set of words associated with the context of a given term. This profile can be based on proximity alone or on some kind of syntactic relation with the term. The profiles are translated from one language to the other using existing bilingual resources. Translated profiles are matched to profiles in the original language using standard word similarity measures from information retrieval. The translations of a given source language term are the terms in the target language with the most similar profiles.

These techniques are very similar to existing monolingual techniques used for automatic thesaurus construction (e.g. Schütze, 1997). Given the nature of the problem and the resources available, many term pairs extracted by this method will not be direct translations, although most will be semantically related. However, finding semantically related terms may be enough to build a good cross-language text retrieval system, and is certainly better than

finding no translations at all. This project will devote significant effort to refining and improving these techniques. So far, they have been tested almost exclusively on general language words. Our challenge will be to extent this work to technical terminology.

2.2 Relation Extraction

Research on automatic extraction of semantic relations from text can be found within different fields. Historically, parsing deals with sub-categorization frames that define the number and position of arguments of predicative verbs and nouns. The semantic complement of this is a set of selection restrictions that help determine the function of each argument of the predicate.

Such *case frames* are used in information extraction to identify entities and their relations automatically from text. For instance, the following case frame, as used by the CRYSTAL system (Soderland *et al.*, 1995,) defines a relation between a symptom and a patient, more in particular, the relation expressed by the verb `denies`:

```
CN-Type: Sign or Symptom
Subtype: Absent
Extract from Direct Object
Active voice verb
Subject constraints:
     Words include “PATIENT”
     Head class:     <Patient or Disabled Group>
Verb constraints:
     Words include “DENIES”
Direct Object constraints
     Head class:     <Sign or Symptom>
```

A similar use of semantic relations is described in the context of information extraction systems like PALKA (Kim and Moldovan, 1993,) AutoSlog (Riloff, 1993; Riloff and Jones, 1999,) AutoSlog-TS (Riloff, 1996,) WHISK (Soderland, 1999,) and Rapier (Califf and Mooney, 1999,) some of which we will describe in more detail below.

Experiments in semantic relation extraction between concepts (terms) from medical texts are based on similar techniques (Oueslati, 1996; Hahn *et al.*, 1999; Craven and Kumlien, 1999; Rindfleisch, 2000.)

Obviously, semantic relation extraction in these systems depends heavily on a robust, but precise assignment of grammatical relations, like subject, (indirect) object, locative adjunct, temporal adjunct, etc. (Buchholz *et al.*, 1999; Yeh, 2000.) If grammatical structure is not taken into account, only a more general relation can be extracted based purely on statistical co-occurrence between terms, see for instance, Swanson and Smalheiser (1997), Maedche and Staab (1999.)

Another prerequisite for accurate semantic relation extraction from text is appropriate semantic tagging of constituents (terms) with a set of (domain specific) semantic classes. For instance, in the example above, the head noun of the subject constituent is tagged with the semantic class `<Patient or Disabled Group>`. Obviously, terms could be assigned

more than one possible semantic class, which would assume some form of sense disambiguation as discussed in Chapter 3.

In summary, the extraction of semantic relations between terms depends on a robust and proper grammatical analysis and semantic classification as mediated by a definition of sub-categorization frames and corresponding selection restrictions (constraints) on arguments and adjuncts.

2.2.1 Information Extraction

The basic task of an information extraction system is to fill one or more domain specific templates with facts extracted from a set of corresponding texts. A template is a data structure with attributes (slots) for each piece of information to be extracted. In order to allow for an automatic extraction of this information, the system needs a domain specific semantic lexicon with entries (case frames) that correspond to the templates to be filled in. One such entry, as used by the CRYSTAL system (Soderland *et al.*, 1995,) was shown above for the concept `<Sign or Symptom>`.

Semantic lexicon entries like these could be defined by hand, but this is rather time consuming. An alternative is to use machine learning techniques that allow automatic generation of such entries from annotated text.

2.2.1.1 AutoSlog (Riloff, 1993)

One of the first systems that were capable of doing this for limited domains is AutoSlog (Riloff, 1993.) Given a set of annotated training texts, AutoSlog proposes appropriate semantic lexicon entries (case frames or *concept nodes*) by applying a set of heuristics that recognize linguistic patterns representing one or more phrases that are likely to be good for activating the entry. If a heuristic successfully identifies the pattern, it generates a conceptual anchor point (a word that should activate the case frame) and a set of enabling conditions to recognize the complete pattern. For instance, consider the following sentence:

(3) `In La Oroya, Junin department, in the central Peruvian mountain range, public buildings were bombed and a car bomb was detonated.`

First, AutoSlog passes the sentence on to a linguistic analysis module, CIRCUS (Lehnert, 1991.) The resulting analysis includes the phrase public buildings were bombed, which matches the pattern `<subject> passive-verb`. This in turn generates bombed as a conceptual anchor point. The resulting semantic lexicon entry is:

```
Name:                target-subject-passive-verb-bombed
Trigger:             bombed
Variable Slots:      (target (subject 1))
Constraints:         (class phys-target subject)
Constant Slots:      (type bombing)
Enabling Conditions: ((passive))
```

Other examples of linguistic patterns used by AutoSlog are:

```
<subject> active-verb      (perpetrator bombed)
<subject> verb infinitive  (perpetrator attempted to
     kill)
passive-verb <dobj>        (killed victim)
gerund <dobj>              (killing victim)
active-verb prep <np>      (killed with instrument)
```

2.2.1.2 AutoSlog-TS (Riloff, 1996)

A major constraint on the portability of AutoSlog is its dependence on annotated domain specific texts, the cost of creating which can be relatively high. Therefore, an extension has been proposed, AutoSlog-TS (Riloff, 1996,) which requires instead of a fully annotated training corpus only a classified set of relevant and irrelevant texts for a particular domain.

AutoSlog-TS operates by generating linguistic extraction patterns (of the sort `<subject> active-verb`) for every noun phrase and then in a second step evaluating each pattern by computing relevance statistics. For example, the sentence `Terrorists bombed the U.S. embassy` might produce two patterns to extract terrorists: `<subject> bombed` and `<subject> bombed embassy`. Relevance statistics computed in the second step will then show whether the shorter pattern is good enough or if the longer pattern will be needed.

Relevance statistics are computed in the following way. An estimate is made of the conditional probability that a text is relevant given that it activates a particular extraction pattern. The motivation for this is that domain specific expressions will appear substantially more often in relevant texts than in irrelevant ones.

Eq 2: Relevance Rate

$$\Pr(\text{relevant text} \mid \text{text contains pattern}) = \frac{\text{rel} - \text{freq}}{\text{total} - \text{freq}}$$

Next, patterns are ordered by use of a ranking function so that a person only needs to review the most highly ranked patterns for acceptance or rejection. The formula promotes patterns that have either a high relevance or a high frequency.

Eq 3: Relevance Ranking

$$\text{relevance rate} \cdot \log(\text{freq})$$

The resulting list of patterns is then used in semantic lexicon construction in ways similar to those described above for AutoSlog.

2.2.1.3 CRYSTAL (Soderland *et al.*, 1995)

The semantic lexicon entries generated by AutoSlog and AutoSlog-TS have fixed constraints on their application. For instance, in the `bombed` example quoted above, the semantic class of the subject must be `phys-target`. If `bombed` occurs with a subject of a different semantic class, this entry will fail. Of course, avoiding erroneous applications of semantic entries is exactly the reason for introducing such constraints (relating to the discussion of selection restrictions above.) Nevertheless, in order to make a more flexible application of semantic lexicon entries possible, in other words to increase recall, constraints could be relaxed without giving up on precision.

This is the approach taken in the CRYSTAL system (Soderland *et al.*, 1995,) already mentioned above. CRYSTAL starts by generating a lexicon entry for each positive instance obtained from an annotated corpus. For example, an entry (CN –concept node- definition) can be generated for `<Sign or Symptom>` from the sentence fragment `Unremarkable with the exception of mild shortness of breath and chronically swollen ankles`, in which `shortness of breath` and `swollen ankles` are annotated with `<Sign or Symptom>`. The entry generated from this instance is:

```
CN-Type: Sign or Symptom
Subtype: Present
Extract from Prep. Phrase “WITH”
Verb = <NULL>
Subject constraints:
     Words include “UNREMARKABLE”
Prep. Phrase constraints:
Preposition = “WITH”
     Words include “THE EXCEPTION OF MILD SHORTNESS OF
     BREATH AND CHRONICALLY SWOLLEN ANKLES”
Modifier class <Sign or Symptom>
Head class <Sign or Symptom>, <Body Location or Region>
```

Obviously, this entry will only apply to the particular instance from which it is generated. In order to make it more widely applicable, some of the constraints need to be relaxed. Semantic constraints are relaxed by moving up the semantic hierarchy or by dropping the constraint. Exact word constraints are relaxed by dropping all but a subsequence of the words or by dropping the constraint.

The CRYSTAL algorithm is as follows:

```
Initialize Dictionary and Training Instances Database
Do until no more initial CN definitions in Dictionary
      D = an initial CN definition removed from the
      Dictionary
      Loop
            D´ = the most similar CN definition to D
            If D´ = NULL, exit loop
            U = the unification of D and D´
            Test the coverage of U in Training Instances
            If the error rate of U > Tolerance, exit loop
            Delete all CN definitions covered by U
            Set D = U
      Add D to the Dictionary
Return the Dictionary
```

CRYSTAL makes useful generalizations by finding similar CN definitions and unifying them by finding the most restrictive constraints that cover both. For example, imagine two definitions with two different class constraints on the subject, one with `<Sign or Symptom>` the other with `<Laboratory or Test Result>`. These can be unified through their common ancestor `<Finding>` in the semantic hierarchy.

In essence, CRYSTAL implements an inductive learning algorithm that tries to cover all positive instances by a minimal set of generalized definitions, avoiding negative instances.

2.2.1.4 Craven and Kumlien (1999)

Another way of looking at information or relation extraction is as a classification task, in which a classifier can be constructed from labeled positive and negative sentences (Craven and Kumlien, 1999.) The method is used in the context of biomedical literature (MEDLINE) to extract relations such as:

```
subcellular-localization (Protein, Subcellular-Structure)
proteins and the subcellular structures in which they are
      found
cell-localization(Protein, Cell-Type)
cell types in which a given protein is found
tissue-localization(Protein, Tissue)
tissue types in which a given protein is found
associated-diseases(Protein, Disease)
diseases with which a given protein is known to have some
      association
drug-interaction(Protein, Pharmacologic-Agent)
pharmacologic agents with which a given protein is known
      to interact
```

The task is to extract instances of a binary relation *r(X, Y).* The first step in this is to identify all instances in a corpus that could possibly express the relation. For this purpose, the words that express the relation are associated (annotated) with semantic classes, using a semantic lexicon. For instance, the semantic class `Subcellular-Structure`, the second argument

in the `subcellular-localization` relation, corresponds to words like `nucleus`, `mitochondrion`, etc. Corpus instances found in this way are sentences, but they could also be larger or smaller text segments like paragraphs or clauses. In this way, the relation extraction task can be framed as a (sentence, paragraph or clause) classification task.

The algorithm is supervised and therefore requires a labeled training set, in which each sentence candidate has been manually labeled as a positive or negative instance. The classification approach uses Naive Bayes, which assumes complete independence between all words in the sentence (bag of words.) Results obtained with this approach are 62% precision at 70% recall, against a baseline of 44% precision at 25% recall. Interestingly, an alternative approach (relational learning) described in the same paper that involves syntactic analysis gives better results on precision (92%), although worse on recall (21%.) Syntactic processing in this approach involves part-of-speech tagging and shallow parsing into noun, verb, or prepositional phrases.

Providing labeled training data for supervised training of the sentence classifier is time-consuming and tedious. Therefore an alternative is sought by exploiting existing resources like biomedical knowledge bases and databases. For instance, an entry for a `subcellular-localization` field in a given biomedical database might contain a reference to the article that established this subcellular-localization fact. Sentences in this article could then be used as weakly labeled instances of the `subcellular-localization` relation. An important point in this is to label as positive instances only those sentences in which both arguments of the relation are mentioned (e.g. the semantic classes `protein` and `subcellular-localization`) and take the rest as negative instances.

In this way, an experiment is set up in automatically, creating and using a weakly labeled training set of instances for the `subcellular-localization` relation. The resulting training set contains significantly more relation instances than the one obtained manually. At the same time, results in classifying are similar and even better than classification based on the manual training set. A Naïve Bayes classifier trained on the automatically acquired training set gives 77% precision at 30% recall.

2.2.2 Grammatical Relation Extraction

Information and relation extraction depend heavily on a robust, but precise assignment of grammatical relations, like subject, (indirect) object, locative adjunct, temporal adjunct, etc. Research in this area builds again on shallow parsing (Abney, 1991; Grefenstette, 1996; Brants and Skut, 1998, among others.) Some recent results in grammatical relation assignment are described in Buchholz *et al.* (1999) and Yeh (2000.)

Using the Penn Treebank II Wall Street Journal corpus (Marcus *et al.*, 1993) as an annotated training set, grammatical relation assignment can be treated as a supervised classification task (Buchholz *et al.*, 1999.) The method used is Memory Based Learning (Daelemans *et al.*, 1998,) in which all training instances are kept in memory in order to take into account also less frequent cases. The approach is similar to the *k*-nearest neighbor, example-based and case-based algorithms (see also the chapter on Word Sense Disambiguation,) in all of which a most likely class hypothesis is constructed on the basis of the set of most similar instances in the training set.

Training instances are constructed according to a number of features that were chosen manually, but extracted automatically for each instance. For instance, the grammatical

relation (NP-SBJ) between Miller and organized in the following sentence can be represented by the following set of 13 features:

```
Not/RB surprisingly/RB ,/, Peter/NNP Miller/NNP ,/,
     who/WP organized/VBD the/DT conference/NN in/NN
     New/NNP York/NNP ,/, does/VBZ not/RB want/VB to/TO
     come/VB to/IN Paris/NNP without/IN bringing/VBG
     his/PRP$ wife/NN .

NP-SBJ : -1 0 0 organized VBD Miller NNP , , who WP
     organized VBD
```

Features 1,2,3 are for distance and intervening VPs and commas, features 4,5 show the verb and its POS, features 6,7,8,9,10,11,12,13 describe the context words. Features are ordered, according to their information gain value (a measure of the reduction of uncertainty about the class to be predicted when knowing the value of the feature.) Most important is the POS of the focus, because this indicates if their could be a relation to the verb at all and if so, what kind of relation (`obj`, `subj`, `loc`, etc.)

By adding chunking information to this, the feature representations become more complex, but also more expressive. Experiments on grammatical relation assignment following this approach show better results, depending on the information (features) taken into account. In fact, the more syntactic structure is added the better precision and recall are: precision increases from 60.7% to 74.4%, recall from 41.3% to 67.9%.

2.3 Text Summarization

Generating an effective summary requires the summarizer to *select*, *evaluate*, *order* and *aggregate* items of information according to their relevance to a particular subject or purpose. There are two kinds of summaries:

Indicative: provides just enough information to the user in order to indicate topic and content, so that he or she can determine whether to read further.

Informative: provides maximal information relative to summary length, so that the summary is meant to be read instead of the much longer original text(s).

The former may be addressed by key-term and key-phrase extraction, but the latter typically requires longer passages such as full sentences or even paragraphs, whether extracted directly from the original text(s) or synthesized from an analysis of said text(s). Most of the work in Summarization has thus far focused on full passage extraction for informative summaries.

Summarization tasks can either be approximated by information retrieval techniques or done in greater depth with fuller natural language processing. One approach is to view summarization as *text-span deletion*, where the system attempts to delete “less important” spans of text from the original document; the text that remains is deemed a summary. The complement view, however, is the most prevalent: locate and extract key passages for inclusion in the summary, as first proposed by Luhn at IBM in the fifties (Luhn, 1958.) Most

of the work in passage extraction applied statistical techniques (frequency analysis, variance analysis, etc.) to finding the most appropriate sentences or paragraphs, but some also addressed other linguistic units such as tokens, names, anaphora, etc. (e.g. Tait, 1983; Paice, 1990; Kupiec *et al.*, 1995; Hovy and Lin, 1997; Mitra *et al.*, 1997; Baldwin and Morton, 1998; Carbonell and Goldstein, 1998.) Other approaches include the utility of discourse structure (Marcu, 1997,) the combination of information extraction and language generation (Klavans and Shaw, 1995; McKeown *et al.*, 1995; Shaw, 1995; Radev and McKeown, 1998, McKeown *et al.*, 1999.) Some work applies machine learning techniques to the task of useful passage identification from free text (Teufel and Moens, 1997; Barzilay and Elhadad, 1997; Strzalkowski *et al.*, 1998.)

Several researchers have extended various aspects of the single-document approaches to address multi-document summarization, typically applied to topically-related document clusters, such as the output from a search engine or a topical similarity clustering process. There are several approaches, including:

- Maximal Marginal Relevance and related methods to reduce potentially-massive redundancy by maximizing objective functions based on information utility in passage selection (Goldstein and Carbonell, 1998; Stein *et al.*, 1999; Goldstein *et al.*, 2000; Radev *et al.*, 2000.)

- Template filling by extracting information – using specialized, domain specific knowledge sources – from the document, and then generating natural language summaries from the templates (Radev and McKeown, 1998.)

- Building activation networks of related lexical items (identity mappings, synonyms, hypernyms, etc.) to extract text spans from the document set (Mani and Bloedern, 1997.)

- Finding co-reference chains in the document set to identify common sections of interest (Baldwin *et al.*, 1999.)

More recently, researchers are addressing multilingual summarization especially with IR-based and statistical methods, as well as genre-oriented methods for producing higher-quality summaries.

3 Cross-language Information Retrieval

3.1 Dictionary-based, Corpus-based and Concept-based Approaches

Cross-Language Information Retrieval (CLIR) is the task of issuing a query in one language and retrieving relevant documents in other languages. It aims to benefit the user in finding and assessing information without being limited by linguistic barriers. The language barrier can be bridged by translating the query, translating the documents, or translating both into an intermediate representation which can be either a pre-defined controlled vocabulary or automatically extracted semantic structures from parallel document collections.

Gerald Salton posed the CLIR challenge as early as 1969, and approached this problem using a hand-assembled bilingual thesaurus in German and English (Salton, 1970.) However, most CLIR work is of more recent vintage; existing approaches fall into the following categories:

1. those based on machine-readable dictionaries or off-the-shell machine translation (MT) systems;

2. those using parallel or comparable corpora to automatically extract domain-specific multilingual thesauri, a latent-semantic interlingua, or trained models for statistical translation; and

3. those employing machine learning techniques for automatic mapping of queries and documents into a pre-defined category taxonomy.

We refer to the above three categories as *dictionary or MT-based*, *corpus-based,* and *concept-based,* respectively. For the MUCHMORE project, we focus on concept-based and corpus-based approaches. Nevertheless, we provide an overview on all three categories below.

3.1.1 Dictionary-based and MT-based CLIR

By *dictionary-based,* we mean using hand-built, general-purpose bilingual dictionaries to translate queries or documents. By *MT-based* we mean using general-purpose (typically rule-based) machine translation systems – we do not mean using statistical machine translation trained on domain-specific document collections. Query translation via machine-readable dictionaries is by far the most common approach in the literature (Grefenstette, 1996; Ballesteros and Croft, 1997; Davis and Ogden, 1997) because of its simplicity. Compared to translating an entire document collection, translating a query by dictionary look-up is far more efficient. However, it is unreliable since short queries do not provide enough context for disambiguation in choosing proper translations of query words, and also because it does not

exploit domain-specific semantic constraints and corpus statistics in solving translation ambiguities. Ballesteros and Croft found that combining dictionary-based approach with query expansion using local context analysis yields better performance(ACL, 1997.) But in spite of such an improvement, dictionary-based methods typically result in significant degradation for CLIR, i.e., 40-80% of the precision and recall of corresponding monolingual retrieval scores.

Using off-the-shell MT systems for query or document translation is also quite popular when such a system is available for the language pairs in consideration. In the Eighth Text REtrieval Conference (TREC-8, 1999) evaluations for CLIR, “at least half of all groups used the SYSTRAN machine translation system in some form for parts of their experiments” (Braschler *et al.*, 1999.) Part of the reason why SYSTRAN is so popular is that it covers all four languages (English, German, Spanish and Italian) included in TREC, and because it is easily accessed through the Internet.

As for the effectiveness of MT-based approaches, the empirical findings so far were rather inconclusive. D. Oard's TREC-6 experiments (Oard, 1998) suggest that the effectiveness may depend on the types of the queries. For short queries (with 1 to 3 words), his results with LOGOS (a commercial MT system) were not better than dictionary-based query translation; for long queries (consisting of a few sentences,) MT-based document translation had better results than MT-based query translation, which was in turn better than dictionary-based query translation. The TREC-7 experiments by Nie *et al*. (1999) partly supported Oard's observation: Translating sentence-based queries, when using SYSTRAN, they obtained better results than those obtained by using corpus-based and dictionary-based methods. However, they did not report parallel experiments with *short* queries for comparison.[1] The TREC results by Gey *et al*. from UC Berkeley are even more interesting (Gey and Jiang, 1999): They found that SYSTRAN outperformed dictionary lookup on the TREC-7 CLIR corpus (containing news stories from the Associated Press and Swiss News Agency), but significantly underperformed dictionary lookup (0.1063 vs 0.2707 in average precision) on the GIRT document collection (available in TREC-8) in the field of social science when using a dictionary automatically extracted from an existing bilingual thesaurus in the same field. Despite the large number of groups who employed SYSTRAN in TREC-8, the best performing systems were the corpus-based MT by Franz *et al*. at IBM and some other corpus-based and dictionary-based methods (Braschler *et al.*, 1999; Franz *et al.*, 1998, 1999.)

It may also worth mentioning that the development cost of rule-based MT system is typically a few person-decades per language pair, and that commercially available MT systems only exist for a few language pairs (typically the most common languages.)

3.1.2 Corpus-based CLIR

Corpus-based learning aims to establish cross-language mappings between queries and relevant documents using empirical associations extracted from bilingually aligned documents. The mapping could be from one natural language (English, for example) to another (Spanish, for example), or from multiple natural languages to an *interlingua* which can either be a pre-defined indexing language or automatically extracted “latent structures”

[1] Another questionable part of the experiments by Nie et al. is that the training documents they used for corpus-based MT were not representative for the test documents.

from data. By learning empirical associations at the lexicon level or the structural level from parallel text, corpus-based approaches exploit domain-specific or application-specific patterns of word usage in context that general-purpose dictionary-based or MT-based approaches do not.

Published work in corpus-based CLIR include the cross-lingual versions of Latent Semantic Indexing (LSI; cf. Dumais *et al.*, 1996,) automatic extraction of cross-language similarity thesauri (Sheridan and Ballerini, 1996; Sheridan *et al.*, 1997; Brown, 1997, 1998,) various forms of Pseudo-Relevance Feedback (PRF; Carbonell *et al.*, 1997; Davis and Ogden, 1997; Ballesteros and Croft, 1997,) the Cross-language Generalized Vector Space Model (GVSM; Yang *et al.*, 1998) and Statistical Machine Translation (StatMT; Franz *et al.*, 1998, 1999; Nie *et al*., 1999.)

In 1997, CMU reported a comprehensive evaluation of the existing corpus-based CLIR methods at that time (including EBT, LSI, GVSM and PRF but not StatMT) where all the methods were implemented and tested under controlled conditions, e.g., with unified stemming, term weighting, similarity measures, etc. (Carbonell *et al.*, 1997; Yang *et al.*, 1998.) With a corpus (namely UNICEF) of Spanish-English parallel documents extracted from the United Nations Multilingual Corpus by the Linguistic Data Consortium (Graff and Finch, 1994,) the corpus-based methods achieved a cross-language performance from 91-101% of their monolingual performance on the same data collection, while dictionary-based query translation only achieved 80% of the corresponding performance of monolingual retrieval.

Evaluations of corpus-based methods on larger document collections were reported in TREC-7 and TREC-8 (http://trec.nist.gov/pubs/trec*/t* proceedings.html). A challenging part of the TREC CLIR evaluations was that the multilingual training documents provided for training are not *parallel* but *comparable* instead. That is, the training documents are topically and chronologically overlapping news stories in different European languages but not semantically identical translations. Such “loosely” aligned training corpora made corpus-based learning a harder problem. Approaches being evaluated include statistical machine translation by Franz *et al.* from IBM T. J. Watson, corpus-based extraction similarity thesauri by Braschler *et al.* from Eurospider, and the combination of n-grams and words by Mayfield *et al.* (the John Hopkins University or JHU.) The IBM statistical MT system was among the one or two top performing systems in the evaluations of TREC-7 and TREC-8 (Franz *et al.*, 1998, 1999.) The JHU method was among the top performing systems in TREC-8 (Braschler *et al.*, 1999.)

Another large CLIR evaluation forum is the NTCIR workshop on Research in Japanese Text Retrieval and Term Recognition (known as “the Japanese TREC.”) NTCIR places more emphasis on Asian languages, currently Japanese and Chinese. It offers a large English-Japanese parallel corpus, consisting of about 180,000 bibliographical records (Kando and Nozue, 1999; cf. the website of the NTCIR Workshop.) Each record contains the title and abstract of an article, plus author-assigned keywords which are also in both languages. The majority of the participants in NTCIR-1 (1999) were Japanese research groups; the best performing system was the one by Gey et al. from UC Berkeley. Their approach (corpus-based) is simple: they obtained a high-quality bilingual lexicon from the author-assigned keywords (phrases in both English and Japanese) to a large collection of articles. This example shows that corpus-based learning may not need to be complex if it uses the proper information. This example also suggests that acquiring quality parallel text that fits the domain or application well is an important part of corpus-based CLIR. Nie at al. illustrated a way to collect bilingual text via Web crawling (Nie *et al.*, 1999.) Resnik reported a similar

effort (Resnik, 1999.) Finding a realistic, systematic and cost-effective way for automated acquisition of bilingual parallel corpora remains an open challenge in corpus-based CLIR.

Scaling corpus-based learning algorithms to very large applications in the real word is another open challenge which has not been thoroughly investigated for many methods. IBM has made significant progress in improving the efficiency of their statistical machine translation algorithms by simplifying their probabilistic models for the purpose of improving retrieval instead of optimizing the quality of translation. As a result, they achieved a translation speed "within an order of magnitude of the indexing time" for the TREC CLIR document collections, that enabled them to translate the documents instead of the queries. This is part of the reason for their good results in the TREC CLIR evaluations because documents offer richer context than queries do for solving ambiguities in machine translation.

3.1.3 Concept-based CLIR

By *concept-based* we mean using a pre-defined category taxonomy as the intermediate indexing language, and using machine learning techniques to transfer queries and documents into the corresponding representations in the indexing language. This approach combines human knowledge (in the sense of using manually developed taxonomy) and corpus-based learning in text categorization (for the mapping from free text to the taxonomy) for CLIR. This is a potentially promising area that has not been well-explored. The CMU team has proposed this unique approach as one of their main foci in the MUCHMORE project.

A superficially similar approach in the literature is the thesaurus-based query translation work by Eichmann *et al*. (1998,) which employs the UMLS Metathesaurus (developed at the National Library of Medicine at the U.S.A.; cf. the corresponding section in Chapter 4.) It is a large hierarchical taxonomy of medical concepts (Medical Subject Headings or MeSH) with extended lexicon entries per concept; some subsets of these concepts have multilingual lexicon entries. The multilingual lexical entries were used to obtain bilingual lexicons for the Spanish-English and French-English language pairs. The approach is essentially dictionary-based query translation, where the dictionaries were derived from the multilingual parts of the UMLS Metathesaurus. They achieved a CLIR performance which was 61-71% of the performance of the corresponding monolingual retrieval on the OHSUMED corpus which is a subset of the MEDLINE database.

Comparing their method to what we referred to as *concept-based*, these two approaches are similar in the first dimension – both use an existing category taxonomy – but fundamentally different in the second dimension: having or not having a learning component for the mapping from multilingual free text (queries and documents) to the controlled indexing language (MeSH concepts in UMLS Metathesaurus, for example.) Eichmann's approach will suffer greatly when the vocabulary coverage of the multilingual entries in the existing thesaurus is small. According to their report, the lexicon entries in Spanish and French parts only mount to 3-5% of the English parts in UMLS Metathesaurus. By common sense, on the other hand, the vocabulary sizes in free-text Spanish and French would be comparable to free-text English. This means that their approach would have a serious problem in translating *arbitrary* Spanish or French documents or queries without losing important information. It is rather puzzling

that, given the 3-5% vocabulary coverage, their CLIR results still achieved 61-71% of the performance in corresponding monolingual retrieval.[2]

The recent work by Gey *et al*. from UC Berkeley in TREC-8 (1999) is even closer to our concept-based CLIR method. For English-German cross-language retrieval over documents (the GIRT collection) in social science, Gey *et al*. extracted a bilingual dictionary from the existing Social Science Thesaurus, which contains controlled terms (categories) in both languages. This part of their approach is similar to what Eichmann *et al.* did with UMLS. The additional part is that they used a k-nearest neighbor classifier to map a query in English to the controlled terms (categories), and used the German version of the controlled terms to expand the German translation (via the bilingual dictionary) of the original English query. The second part of their method is similar to (but not the same as) the query-to-category mapping part in our proposed concept-based CLIR. What our method offers in addition is the learning component for the document-to-category mapping, and a mechanism for matching or ranking the categorized documents with respect to the categorized query. In other words, their method will suffer significantly when there is a large gap between the document vocabularies and the controlled vocabulary – which is a well-known phenomena in practice. Our method addresses this problem by training separate classifiers, one for query transformation in the query language, and another for document transformation in the document language, both into a common concept vocabulary.

To be more precise, the concept-based approach proposed by CMU for MUCHMORE bridges the vocabulary gaps between free text in multiple languages and the controlled indexing language (MeSH) through supervised classification techniques, including k-Nearest Neighbor, Linear Least Squares Mapping, Support Vector Machines, and so forth. A set of manually categorized documents and/or queries from the same domain of the application in consideration will be chosen as the training sets to obtain empirical associations between multilingual vocabularies and categories in the MeSH taxonomy. There is a rich literature in both medical informatics and text categorization studying and assessing the effectiveness of different techniques to solve this problem (Yang and Chute, 1993, 1994a, 1994b; Yang, 1999.) By combining the strengths of corpus-based learning and knowledge-based generalization via concepts, we hypothesize that the concept-based approach will not only be effective for automatically bridging the language barrier in CLIR, but also beneficial for supporting concept-based browsing by the user. The latter point has started to gain research attention in monolingual retrieval (Chen and Dumais, 2000; Dumais and Chen, 2000,) but its benefit in CLIR has yet to be investigated. Such an investigation is one of the aims in the MUCHMORE project.

[2] The experiments by Eichmann et al. were conducted by translating Spanish and French queries into English and retrieving documents in English. The Spanish and French queries are their manual translations of a set of 106 English queries; whether the Spanish and French words in the queries were carefully chosen to leverage the limited UMLS multilingual vocabulary was not reported.

3.2 State of the Art in Parallel-Corpus Based Methods

3.2.1 State of the Art in Parallel-Corpora Acquisition

One problem with corpus-based information-retrieval and translation methods is that one needs a (usually large) parallel corpus, which may be difficult to acquire for the language pair of interest. The explosion of information on the World Wide Web yields an interesting new source of parallel text, which a few researchers have been harnessing through automated retrieval from the web. Most projects requiring parallel text, however, still rely on existing corpora such as the Hansards and other parliamentary proceedings, or create their own by hiring translators (LDC, 1997; Graff and Finch, 1994.)

The most prominent results on parallel-corpora acquisition to date are those of Phil Resnik at the University of Maryland (Resnik, 1998.) Resnik's approach is to use a web spider to collect pages containing certain key expressions which tend to indicate that a certain link points at a translated version of the page, and then filter the retrieved pairs of pages by ensuring structural parallelism of the HTML tags within the pages.

3.2.2 State of the Art in Corpus-Based Cross-Language Information Retrieval

Just as the explosion of information on the Internet has enabled the automated acquisition of parallel corpora, it not only enables but also necessitates cross-language information retrieval (Hovy *et al.*, 1999.)

One of the main approaches to cross-language retrieval has translating of either the query or the document collection. With the availability of substantial parallel corpora, one obvious method of generating translations is to use a corpus-based machine translation system. The state of the art in corpus-based machine translation approaches is reviewed in the next section. Other approaches which do not involve explicit translation include the Generalized Vector Space Model, bilingual Latent Semantic Indexing, and bilingual pseudo-relevance feedback (Yang *et al.*, 1997, 1998.)

3.2.3 State of the Art in Corpus-Based Translation

Over the past 10 to 15 years, the increased availability of computational power, memory, storage, and parallel texts has enabled vigorous activity in the field of corpus-based translation – using already-translated texts as the basis for translating new texts. Corpus-Based Machine Translation (CBMT) can be subdivided into several categories, although the boundaries between categories are fluid: translation memory example-based machine translation, and statistical machine translation (the term Memory-Based Translation is also used, sometimes for translation memories and sometimes for example-based systems.) There are also hybrid systems which use multiple CBMT approaches or combine CBMT with traditional rule-based machine translation (IAI 36, 2000.)

Translation memories simply store prior reference translations of text and retrieve the nearest match in their database. It is then up to the human translator to clean up the retrieved translation and to account for the differences between the entry in the database and the actual passage to be translated. Even this very simple technology proves to result in a major increase in (human) translator productivity, and is commercially available in a number of products, such as Atril's Deja Vu, IBM's Translation Manager (TM2), SDL International's SDLX (www.sdlintl.com), the Trados Translator's Workbench (TWB; Heyn, 1996,) Star's Transit, and Zeres GmbH's ZERESTRANS (Zeres, 1997.)

Example-based translation systems use a corpus of pre-translated example phrases and sentences as a basis for translating previously unseen text. Approaches range from the purely lexical (string matching) to fuzzy matches between parse trees (Brown, 1996, 1999, 2000; Carl, 1999; Collins, 1999; Cranias *et al.*, 1994; Furuse and Iida, 1992; Gvenir and Cicekli, 1998; Maruyama and Watanabe, 1992; Nagao, 1984, 1985; Nirenburg *et al.*, 1994; Sato, 1991; Sato and Nagao, 1990; Sumita and Iida, 1991; Veale and Way, 1997.)

Statistical machine translation systems use the parallel training corpus to build a probabilistic model of the possible translations for words and the reordering of words between languages. After training, the corpus is no longer required, as all translation is performed using the statistical model (Brown *et al.*, 1990, 1993; Al Onaizan *et al.*, 1999.)

Among the less-common CBMT approaches are the use of neural networks for EBMT (McLean, 1992.) Various hybrid corpus-based approaches have also been tried, such as translation memory plus EBMT (Carl and Hansen, 1999.)

In addition to purely corpus-based methods, there has been considerable recent activity in hybrid systems which augment traditional rule-based translation (RBMT) using some form of corpus-derived knowledge. These include adding statistical MT to RBMT (Rayner and Bouillon, 1995; Streiter *et al.*, 1999; Choi *et al.*, 1998; Jung *et al.*, 1998) and adding EBMT to a rule-based system (Carl *et al.*, 1999b.)

Finally, some researchers have added rule-based methods to otherwise corpus-based systems, e.g. rule-based added to statistical (Chen and Chen, 1995.)

Parallel corpora are also mined for the implicit knowledge contained within them, typically terminology and bilingual dictionaries/collocations (Catizone *et al.*, 1993; Brown, 1997, 2000; Melamed, 1996; Tanaka, 1996; Tiedemann, 1998a, 1998b.)

3.3 Non-parallel (Comparable) Corpora

3.3.1 Non-parallel vs. Parallel Corpora

Non-parallel corpora are intended to be a cheap alternative for parallel corpora. Since parallel corpora contain real, manually produced translations, they are rare, and their availability is usually limited to certain fields, like government data from multilingual countries. The

production of additional parallel corpora is very expensive, especially for the sizes that are necessary to do meaningful work in MT, NLP and IR.

There are several degrees of non-parallelism that can be considered as an alternative, for example corpora containing near translations, corpora containing similar items, corpora with items that cover at least the same domain, or corpora that have subcollections with no similarity at all. Corpora with items that have at least some degree of similarity are usually referred to as *comparable corpora*.

Typically, the more similarity, the more interesting the corpus is for processing. However, there are areas where even completely non-parallel corpora can be useful, such as demonstrated in CLIR work by (Ballesteros and Croft, 1998). In this case, the corpora are consulted prior to and after the translation step to help with disambiguation of terms.

3.3.2 Alignment Granularity

Many procedures working on non-parallel corpora (and also on parallel corpora) need some form of text alignment as a prerequisite. Parts of different texts are matched, thus relating them. In the multilingual case, this has to be done across languages. Depending on the similarity of the items in the comparable corpus, different levels of *alignment granularity* are usually chosen. Parallel and near-parallel corpora are usually aligned on the sentence or even the word level (Gale and Church, 1993), which is a necessity for many NLP processes. Corpora with less closely related items can instead be aligned on the paragraph (Franz *et al*, 1998) or even just the document level (Braschler and Schäuble, 1998). Such alignments can still be used for various corpus-based approaches in CLIR, for example to train a statistical machine translation model (Franz *et al*, 1998) or a similarity thesaurus (Schäuble and Knaus, 1992, Qiu and Frei, 1993, Sheridan *et al.*, 1997).

3.3.3 Use of Non-parallel Corpora for CLIR

Non-parallel corpora can be exploited in various approaches to CLIR:

- PRF (Pseudo Relevance Feedback) (Braschler and Schäuble, 1998)
- GSVM (Generalized Vector Space Model) (Carbonell *et al.*, 1997)
- LSI (Latent Semantic Indexing) (Landauer and Littman, 1990)
- Similarity Thesaurus (Sheridan and Ballerini, 1996)
- Word sense disambiguation after dictionary lookup (Davis, 1996)

All of these approaches can work with non-parallel corpora that are aligned at the document level. Alternatively, it would seem feasible to use alignments on a paragraph level.

3.3.4 Producing Non-parallel Corpora

Non-parallel corpora can either be formed through coupling of sufficiently similar document sources (say, combining documents from the English Associated Press newswire with the French Agence France Presse) or through mining techniques on resources like the World Wide Web (Nie *et al*., 1999).

To align similar document sources, items from one collection potentially have to be compared to all other items in the other collection. Since this is usually not practical, restrictions like the date of the items or rough classifications are used to prune the search space (Sheridan *et al*., 1998, Braschler and Schäuble, 1998).

Building comparable corpora from the Web implies acquiring of a sufficient amount of web data and analyzing it for indications that the page is also available in a translated form. Such an indication can exist in the form of a phrase (e.g. click here for a French version of this document) or a URL pattern (e.g. http://some.host/english/...).

3.3.5 Producing Alignments

Several approaches exist to align non-parallel corpora. Possibilities range from sophisticated linguistic analysis to statistical methods based entirely on word matching, frequency analysis and length analysis of the texts (Gale and Church, 1993, Chen, 1993, Braschler and Schäuble, 1998.) Much work has been done on the task of aligning specific classes of terms, such as terms from a technical domain (Fung and McKeown 1997). While not directly interesting for CLIR, such approaches can be used for further alignment iterations (Fung and McKeown, 1994). If the algorithm works on the word-level for general domains, it can be interesting to extract bilingual lexica for further CLIR use (Fung, 1995).

3.3.6 Examples of Non-Parallel Corpora for CLIR Evaluation

With the arrival of forums for evaluation of CLIR systems, the first comparable corpora for CLIR complete with relevance assessments have been created. The TREC CLIR track produced a quadrilingual corpus containing mostly news wire articles in English, French, German and Italian (Braschler *et al*, 1999). Its successor, the newly formed CLEF initiative, has published a first version of a corpus for the same languages, containing mostly newspaper texts. TREC now has an Asian language CLIR track, experimenting with Chinese newspaper texts, but the corpus unfortunately is monolingual at this time. NACSIS is building a Japanese/English test collection for use with the NTCIR retrieval evaluation workshops. Only some of the documents have English translations.

3.4 Evaluation

3.4.1 IR Evaluation

Evaluation has traditionally been a major research focus for the IR community. Research on evaluation of retrieval systems dates as far back as the mid-1950s. Important early work includes the Cranfield tests (Cleverdon and Mills, 1963) and the MEDLARS evaluation (Lancaster, 1969). In the Nineties, large scale evaluation forums were pioneered by the TREC conference series (Harman, 1995), which are now in their ninth year. Much of this work concentrates on the question of the quality of the search results, i.e., does the system successfully retrieve items relevant to the query while rejecting irrelevant items? This, the quality of the results, or more generally, the *effectiveness* of the system, however, is clearly only one aspect of system evaluation. A wide range of other system characteristics could conceivably be considered, from low-level performance issues to questions regarding the user interface/interaction.

3.4.2 Aspects of Evaluation of IR Systems

3.4.2.1 Effectiveness

As mentioned, a major part of the effort traditionally spent on research into IR evaluation is centered on questions of effectiveness. Therefore, the methodology for automated testing is well-developed and documented. A number of evaluation forums/conferences have been set up that allow cross-system comparisons. Evaluation of system effectiveness will also be a main focus in MUCHMORE.

3.4.2.2 Efficiency and Acceptability

Performance-, or efficiency-related questions are not so much a focus of a project such as MUCHMORE, which concentrates on developing Prototype software. Acceptability is addressed by the user requirements report and will be assessed by measuring the project results against the user needs laid out in this report.

3.4.2.3 Notion of Relevance

Central to nearly all effectiveness testing of IR systems is the notion of *relevance* of a document with regard to a query by the user. To calculate effectiveness measures, a document is regarded as relevant if it contains information that meets the user's information need ("the answer to the question"). Otherwise, the document is considered irrelevant. The goal of all the measures in some form or other is to retrieve a maximum number of relevant documents without also retrieving an excessive amount of irrelevant items.

A potential major problem with this approach is that clearly, relevance in this sense is very subjective. What one person perceives as relevant can be regarded as completely irrelevant by an onlooker with a different background. Even the same person can change his or her judgment of relevance over time. Additional factors like novelty of the retrieved information, or the user's ability to understand it, can also play roles. A comprehensive overview of the notion of relevance and its aspects can be found in (Mizzaro, 1998). A review of Mizzaros ideas can be found in (Draper, 1998).

It is therefore important that the effectiveness measure work reliably in face of modest changes to the set of relevance judgments that is used for its computation. Research on this question in large scale evaluation environments is relatively new. Groundbreaking in this regard was the study by Voorhees (1998), which has recently been followed up with further investigations (Buckley and Voorhees, 2000).

3.4.2.4 Measures

3.4.2.4.1 Recall/Precision

Recall and Precision are the most commonly used measures for effectiveness. They are defined as follows:

Eq 4: Precision and Recall

$$\text{Precision} = \frac{\text{number of relevant documents retrieved}}{\text{number of documents retrieved}}$$

$$\text{Recall} = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents}}$$

Since the result lists of probabilistic search engines are typically very long (essentially, every document in the collection is ranked), these measures are usually calculated at various levels, i.e. Precision at *n* (Precision of the set of the top *n* documents retrieved) or Precision at Recall *x* (Precision after *x* percent of all relevant documents are retrieved). For obvious reasons, it is popular to express system performance with a single number. While this may be convenient for comparison purposes, it is also dangerous, because an oversimplification of system differences can hardly be avoided. Usually, the so-called "average precision" is the most popular "one figure measurement." It is computed as the average of a set of precision values at different recall levels (using 11 recall levels from 0.0 to 1.0 in 0.1 increments is a popular choice). An in depth discussion of the merits of this and similar approaches is given in (Hull, 1996).

3.4.2.4.2 Other measures for effectiveness

An alternative to Recall and Precision are utility measures. In this case, "bonus points" are added to an imaginary score if a relevant document is retrieved, whereas "penalty points" are deducted from the score if an irrelevant document finds its way into the retrieval result. The higher the resulting score, the better presumably the system performs. This way of evaluation is suited very well for filtering experiments: the system automatically sends documents of potential relevance to a user who inspects them. The more relevant information the user is sent, the bigger his or her satisfaction. If, however, receiving more relevant items comes at the expense of also having to sift through more noise, the user satisfaction can be assumed to suffer. These assumptions can easily be modeled with the bonus and penalty points. Therefore, utility has been used with success in evaluation of filtering experiments (see eg. Hull, 1999). The methodology is less suited for ranked lists from "query-answer" retrieval experiments, since it does not incorporate the ranking information into the score calculation (i.e., whether a relevant document is found at the top of the list or at the bottom of the list does not influence its contribution to the overall score).

A further effectiveness measure is *overlap*. For use of this measure, it is assumed that a perfect or "very good" result is already available. Then various ratios of overlap between the known, proven results and the new experimental results can be calculated. Possible application fields include retrieval on OCR-texts: if a "perfect" transcript is available, then the quality of retrieval on the OCR-derived texts can be measured by calculating the overlap with the results from retrieval on the transcript. Clearly, this strategy is also applicable to CLIR: a monolingual experiment serves as the "good baseline" against which the cross-language results are measured. If a test collection with relevance assessments is available, however, this approach is considered inferior and usually avoided (Carbonell *et al.*, 1997).

3.4.2.5 Test collections

The popular measures of recall and precision, and many more effectiveness measures, build on the notion of relevance. This means that for their calculation, some form of "relevance assessment" is needed. While for precision there seems to be a straightforward way (evaluating the top *n* documents of the result) for calculation, it is much harder to determine recall. Theoretically, it is necessary to review every document in the collection to determine an exact figure for recall. Unfortunately, today's document collections are likely to be at least several magnitudes too big for such an endeavor. Therefore, early test collections, for which recall was calculated using this method, seem ridiculously small from today's viewpoint. In order to create bigger test collections, recall is therefore usually estimated with the help of the so-called pooling technique (Sparck Jones and van Rijsbergen, 1975). The assumption is that if a sufficient number of adequately different systems is used, it is unlikely that too many relevant items can go undetected. Thus, by pooling all relevant documents retrieved by all these systems, a good estimate of the true number of relevant documents should be obtained. Because this definition is rather vague (what is a sufficient number of systems and how different do the individual systems need to be?) evaluation forums were created that aim to bring a maximum number of participants from different research groups together. By running identical queries on their systems, these groups can help build the pools and thereby ultimately the test collection. This style of evaluation was pioneered by the TREC conferences, which were part of the Tipster program (Harman, 1993; Voorhees and Harman, 2000).

3.4.3 Interactive vs. Non-Interactive Experiments

A major criticism with a lot of the work mentioned so far is that it is thoroughly automated, in "batch processing" style. There are many questions surrounding the interaction of the user with the system that are left unaddressed: the use of feedback mechanisms the user may have available, the influence of result presentation on the usefulness of the results to the user, the usability of the system and many more. A major reason for the relative neglect of these questions for a long time may be the huge effort needed to conduct a meaningful experiment dealing with these questions. TREC introduced an interactive track to investigate some of these issues beginning with TREC-5 (Harman, 1997; Over, 1997). Unfortunately, interactive experiments on a large scale are very difficult and expensive to set up and need to reach a quite substantial scale if any statistically meaningful differences are to be detected.

3.4.4 Cross-Language track at TREC

Over the years, the TREC series of conferences began to look at more and more "specialized" aspects regarding IR system evaluation. Among the so-called *tracks* included were "interactive", "confusion", "database merging", "filtering" and many more. Most importantly from the viewpoint of our project, tracks addressing languages other than English were introduced starting with TREC3 in the form of a (monolingual) Spanish track. Later, a monolingual Chinese track was also introduced. While these tracks disappeared after TREC5 and TREC6, respectively, they were precursors for the introduction of a cross-language information retrieval track at TREC. Starting with TREC6, such a CLIR track was offered to the TREC participants. Initially, the evaluation was limited to bilingual retrieval with a choice of English, French or Italian for topic and document language (some unofficial additional topic languages were also offered) (Schäuble and Sheridan, 1998). With TREC7, this was expanded to include "real" multilingual retrieval, i.e. searching on a collection containing documents in many languages (Braschler *et al*., 1999, 2000). Cross-Language retrieval evaluation added some new problems to the evaluation task such as the need for a translation of the test queries (they are provided to the participants in all languages, so they have a free choice) and the problem of assessing documents in different languages, meaning that assessors competent in all these languages have to be found. It also means that more than one assessor is working on the same query. These and more issues are discussed in (Braschler *et al*., 2000). The result of the TREC CLIR tracks was the first large cross-language test collection.

In 1999, the first NTCIR (Kando *et al.*, 1999) workshop was held in Japan. NTCIR is an evaluation forum modeled on many of the ideas that were pioneered in TREC. It also offered bilingual retrieval in English and Japanese.

In 2000, the CLIR track was separated from TREC and expanded to a new standalone forum, CLEF (Cross-Language Evaluation Forum). While NIST, the backer of TREC, remains involved, CLEF is organized in Europe with funding from the European Commission. The spin-off allowed to significantly expand the activities, which lead to increased participation (20 groups for CLEF-1) and which will help to produce better test collections. It is likely that CLEF, as it expands, will eventually also spawn specialized tracks, e.g. for audio CLIR, very much in the way TREC did for monolingual retrieval. TREC retains a very limited CLIR track specifically for Asian languages in TREC-9. TREC no longer offers CLIR evaluation on "Western" languages.

3.4.5 Other notable Non-English and Cross-Language Evaluations

3.4.5.1 Amaryllis

Launched in 1995, Amaryllis investigates monolingual information retrieval in French (Coret *et al.,* 1997). Amaryllis has completed their "second cycle" in 1999. There was a limited cross-language track, using a very small parallel corpus, and allowing bilingual experiments only. The languages covered were English, French, German, Italian, Portuguese and Spanish (Chaudiron and Schmitt, 2000).

More information on the Amaryllis project can be found online at http://www.inist.fr/accueil/profran.htm.

3.4.5.2 IREX

IREX is an initiative for the evaluation of Japanese monolingual retrieval. The first IREX campaign has ended with a workshop held jointly with NTCIR in 1999 (Sekine and Isahara, 2000).

More information on the IREX project can be found online at http://www.cs.nyu.edu/cs/projects/proteus/irex/index-e.html.

4 Word Sense Disambiguation

4.1 Overview

Words mostly have more than one interpretation, or *sense*. If natural language were completely unambiguous, there would be a one-to-one relationship between words and senses. In fact, things are much more complicated, because for most words not even a fixed number of senses can be given. Therefore, only in certain circumstances and depending on what we mean exactly with *sense*, can we give restricted solutions to the problem of *Word Sense Disambiguation* (WSD.)

4.1.1 Word Sense Disambiguation

Word sense disambiguation is a task relative to what we mean with *sense*. Lexical semantic ambiguity can be between *homonymic* senses (`the river bank` vs. `the money bank`), *systematically polysemous* senses (the related `animal` and `food` interpretations of `rabbit`), *vague* senses (`give me your - left` or `right - hand`) and everything in between.

The vagueness of hand, without much context to conclude if the left or right hand is being indicated, traditionally is not seen as a disambiguation task. Still, we do know that a hand can be further specified as being the right or left hand. If this information is additionally given, the representation will be updated accordingly.

We see a similar process also in the case of systematic polysemy, as with the meaning of `rabbit` in the context of:

```
(4)  The rabbit I shot yesterday was delicious.
```

In this case, the information expressed is not only on the `food` interpretation, but also on the `animal` interpretation in connection to the food that is made from it. Compare the following context, in which the `food` interpretation could be inferred, but is not explicitly referred to:

```
(5)  I shot a rabbit yesterday.
```

With homonyms (also known as *homographs*), there is rather a choice between two or more competing interpretations, which corresponds to the traditional view of lexical semantic ambiguity. If one interpretation in processing is chosen, the other(s) drop(s) out of the semantic representation completely. Here, the domain of application may be a decisive factor in disambiguation. Homonyms may be unambiguous in certain domains. Obviously, `bank` is likely to mean `money bank` in the financial domain.

4.1.2 Methods

WSD involves two parts, a semantic lexicon that associates words (types) with sets of possible senses and a method of associating (annotating, tagging) occurrences of these words (tokens) with one - or more, in the case of systematic polysemy - of its senses. The systems and algorithms that have been developed for this cover the full spectrum of methods developed in Natural Language Processing (NLP) and Artificial Intelligence (AI) in general. For the purposes of this report we can group these as follows:

4.1.2.1 Knowledge-based

The construction of the tag set (the senses used and their association with word types in the semantic lexicon) and the tagging (disambiguation between possible senses and association of the preferred sense with a given word token) are both supervised.

These approaches use small, but deep, handcrafted lexicons to analyze a small number of examples in a non-robust way, that is, the systems can handle only certain input. Generally, the choice between alternative interpretations is based on some measure of similarity between words and their contexts, based on semantic networks, built-in-preference orderings or software agents ("word experts") running in parallel (Small, 1980, 1983; Hirst, 1998; Adriaens and Small, 1988.) All of them, however, rely on pre-coded, domain-specific knowledge, which is the heaviest cost factor in work on WSD, yet indispensable (the "knowledge acquisition bottleneck.") Typically, handcrafted rules are constructed, according to given examples. There is no automatic training of the system.

4.1.2.2 Hybrid: Knowledge-based/Empirical

The construction of the tag set is supervised, but the training for the tagging can be either supervised or unsupervised. In the first case, corpora that have been manually annotated are used for training. In the second case, corpora that have not been annotated are used for training.

These approaches combine hand-crafted knowledge bases with empirical data derived from large corpora. These approaches use large scale, more shallow, hand-crafted lexicons (WordNet, Roget, lexical database versions of LDOCE, OALD, etc.) to analyze text in a robust way, that is, the systems can handle free, naturally occurring text.

Most of these systems became possible thanks to technological advances and the availability of large scale knowledge bases, such as machine-readable dictionaries (Lesk, 1986), thesauri such as Roget's (Yarowsky, 1992), and computational dictionaries like WordNet (Resnik,

1995) (Ng and Lee, 1996.) However, each is still plagued with specific problems, mostly stemming from limited suitability of the knowledge base used for the task at hand. Also, the knowledge acquisition bottleneck arises again, since for proper training of a system based on corpora, semantically annotated corpora of useful proportions would be needed. Although there are a few such efforts, the resulting corpora are small. A good compromise seems to be bootstrapping from a small pre-tagged subset of words (Yarowsky, 1992.)

4.1.2.3 Empirical

The construction of the tag set and the training for the tagging are both unsupervised. These approaches use no external knowledge base at all, but instead derive the tag set itself from the corpus as well. This might be called "self-organizing" WSD. It seeks to do without the pre-defined set of alternative senses, inferring them instead by working, as it were, in the opposite direction: A corpus is used to classify words based solely on patterns of occurrence. The resulting clusters are presumed to represent senses. The two stages that this process consists of are (1) clustering the occurrences of a word into a number of categories and (2) assigning a sense to each category. This idea was first discussed in Schütze (1992). It can dispense with the sense labeling step if the results are only used machine-internally. To stress this subtle but important difference, the method is called W*ord Sense Discrimination* in Schütze (1998). There are still some problems with it, most notably the close dependence of the resulting classification on the training corpus and the choice of clustering granularity.

4.1.3 Evaluation

A problem with the various methods proposed for WSD is the lack of a standardized evaluation metric. Publications often focus on only a few words (in a recent special issue on WSD of Computational Linguistics, two of four article concentrate exclusively on the three words `line`, `serve`, and `hard`, while many other papers use some other small set of words.) In hand-tagged corpora, disagreement between human judges invariably introduces considerable noise (Veronis, 1998.) Moreover, it is difficult to draw comparisons across domains. These problems motivated the SENSEVAL project, aimed at developing a standardized evaluation metric for WSD systems and holding regular tournaments. We will discuss SENSEVAL in Section 4.6.2 below.

4.1.4 Cross-Linguality

From a cross-lingual point of view, word sense disambiguation is nothing more than determining the appropriate translation of a word (or lexical item in general.) Therefore, translation always presupposes word sense disambiguation, although not necessarily in any explicit form. That is, in order to translate a word from language A into language B, we only need to know if the word in language B expresses the same meaning as the word in language A. We do not necessarily need to know what exactly that meaning is. In recent, corpus-based approaches to cross-lingual processing, word sense disambiguation is therefore left implicit (Resnik and Yarowsky, 1997.) At the same time, however, such automatically extracted translations can be used to tune existing lexical semantic resources (Ide, 1999.)

4.2 Knowledge Based Approaches

Traditional approaches to the resolution of lexical semantic ambiguity involve deep representations of knowledge on the semantic context of words. This includes constraints on the arguments of predicates and frame representations to capture associative relations between concepts in general.

4.2.1 Selection Restrictions

Linguistic theory has long argued for restrictions to be imposed on the semantic type of arguments. These selection restrictions are needed to resolve the form of lexical semantic ambiguity we find in sentences like:

```
(6)   John went home.
(7)   John went conservative.
```

In the first sentence, the verb went expresses movement, whereas in the second sentence it is the figurative interpretation of this verb that is being expressed. The selection restrictions imposed by went on the second argument (direct object) will be something like location with the first and psychological with the second interpretation.

Selection restrictions can be thought of as a set of patterns that are matched against syntactic structures. The patterns are definitions of allowed semantic structures built up by predicates and their arguments. To make full use of this, a semantic lexicon needs to be build that has semantic types for all words. By classifying the types in a hierarchical order, we get a type hierarchy that can be used with subsumption in semantic processing.

It is these kinds of considerations that have lead to construction of WordNet (Miller, 1995) and similar lexical semantic resources (EuroWordNet: Vossen, 1998) Nevertheless, although WordNet organizes words in semantic types (or rather synsets - sets of synonyms), selection restrictions on predicates are not represented. Recent initiatives on extensions of WordNet, like VerbNet and FrameNet (Baker *et al.*, 1998), are set up to remedy this.

4.2.2 Marker Passing

A type hierarchy will be an essential component of a more general semantic network that represents words by interconnected concepts. Each concept may be represented by a frame, in which associations with other concepts are captured through frame slots and fillers for these slots. For instance the following information could be available in a frame for a concept corresponding to the word library (The example is in the frame-based language Frail (Hirst, 1988)):

```
[     frame :     library
      isa   :     institution
      slots :     (function  (store-books    lend-books))
                  (employee  (librarian))
      ... ]
```

Frames and similar representations of associative relations between words / concepts can be used to resolve lexical semantic ambiguities in probably very much the same way as humans would. For instance, in the following sentence, the meaning of `tree` will be determined by its association with `linguist`.

```
(8)  The linguist drew a tree.
```

In a computational system, the association between these two words needs to be computed by measuring the distance between the concepts expressed by them in the semantic network. This technique, *marker passing* (Charniak, 1981; Hirst, 1988), is inspired by the idea of spreading activation between nerve cells in psycholinguistic models: “Marker passing can be thought of as passing tags or markers along the arcs of the knowledge base, from frame to frame, from slot to filler ... It is a discrete computational analogue of the spreading activation models often used in psychological models of memory ...” (Hirst, 1988.)

Both selection restrictions and marker passing make use of the fact that words interact semantically. “In one sense, every word in a sentence interacts semantically with every other word, and also with words in neighboring sentences. But we must distinguish between a type of interaction which is precisely regulated by the syntactic structure of the sentence, and a more diffuse type of interaction, not dependent on syntax, but merely on discourse propinquity ...” (Cruse, 1986.)

As laid out in the next sections, in more recent approaches to lexical semantic ambiguity resolution (now commonly referred to as word sense disambiguation) attention moved away from the “precisely regulated” to the “more diffuse” types of interaction.

4.3 Hybrid Approaches: Using Knowledge Bases with Corpora

4.3.1 General Remarks

As we mentioned in the introduction, the possible senses of a given word may be provided to the algorithm externally. In such a setting, the task of WSD can be regarded as a *classification* task. (The alternative situation, in which senses are *discovered* from an underlying corpus automatically, will be discussed below in Section 4.4 below.) The classification algorithm is used to decide for each occurrence of the word to which of the senses it belongs, based on the context surrounding it. Several issues are involved in this: supervised or unsupervised training, semantic annotation, and context.

4.3.1.1 Supervised and Unsupervised Training

Training of the resulting classifier can be either supervised or unsupervised. Supervised training assumes a manually annotated corpus of training examples, in which each word occurrence is labeled with one (or more) of available senses. With unsupervised training, no such annotated corpus is available. Instead, a classifier needs to be build based on co-occurrence constraints given all available senses.

Depending on supervised or unsupervised training, different algorithms could be used. Supervised methods include (naïve) Bayesian modelling (Gale *et al*., 1992), decision lists (Yarowsky, 1994) and exemplar-based approaches -- k-nearest neighbor / case based reasoning (Ng, 1997), among others (a comparison of methods in Mooney, 1996.) Unsupervised methods include the algorithms proposed by (Yarowsky, 1992), (Resnik, 1997) and (Agirre and Rigau, 1996.)

4.3.1.2 Semantic Annotation

Supervised methods in Word Sense Disambiguation use semantically annotated corpora to train machine learning algorithms in deciding which word sense to choose in which contexts (Ng and Lee, 1996; Ng, 1997, Wiebe *et al*., 1997.) In such corpora, words (or lexical items in general) have been tagged manually with a semantic class, as given by a particular lexical semantic resource, e.g.: WordNet, Roget, LDOCE, etc. (Kilgarriff, 1998.)

A problematic issue in (semantic) annotation is the agreement between human annotators (Inter-Annotator Agreement: IAA.) This arises from the fact that semantic annotation is not a clear defined task by itself. In fact, IAA depends largely on the semantic resource available (the level of fine- vs. coarse-grainedness of sense distinctions) and on the lexicographic skills (including domain expert knowledge) of the annotators.

Some studies have been done in this direction (Fellbaum, 1997; Ng, 1999), but to make WSD more efficient IAA needs to be studied in even more detail, in order to get a better understanding of which senses should be distinguished at all (given agreement in human judgement) and for which purpose (domain expertise of annotators.)

4.3.1.3 Context

Context is the single most important factor in WSD. It can reach from the very local to the very global, all of which is needed in useful combinations, in order to determine word meaning in a robust and precise manner. Local context could involve windows of surrounding words, local collocations, syntactic relations, part-of-speech and morphology. Global context could build on semantic distinctions in discourse sections (Gale *et al*., 1992), or over domains (Turcato *et al*., 2000; Buitelaar, 2000; Sacaleanu, forthcoming).

4.3.2 Unsupervised Methods

One of the first steps from the pure knowledge based approaches described earlier towards the corpus based approaches described in the remainder of this report, was the use of machine

readable dictionaries (MRD's) as large lexical knowledge bases (LKB's) derived from the print tapes normally used by printers. LKB's have been derived from dictionaries such as Webster's (Amsler, 1980?), OALD (Lesk, 1986) and LDOCE (Boguraev and Briscoe, 1989; Wilks *et al.*, 1996; Copestake, 1992.)

4.3.2.1 Lesk 1986

The method for using such an LKB for WSD as described in (Lesk, 1986) is representative for most of these approaches. The basic idea is to disambiguate between senses of a word (token) by comparing its dictionary definitions with those of the other words in its context. Compare, for instance, the dictionary definitions of the words `ash` and `coal` in Webster's:

ash
1. the **solid** residue left when **combustible** material is thoroughly **burned** or is oxidized by chemical means
2. ruins
3. the remains of the dead human body after cremation or disintegration
4. something that symbolizes grief, repentance, or humiliation
5. deathly pallor

coal
1. a piece of glowing carbon or charred wood: ember
2. charcoal
3. a black or brownish black **solid combustible** substance formed by the partial decomposition of vegetable matter without free access of air and under the influence of moisture and often increased pressure and temperature that is widely used as a natural fuel; pieces of a quantity of the fuel broken up for **burning**

When these words co-occur in a text they can be mutually disambiguated by computing the number of matching words in their definitions. In this case, the overlapping words are `solid`, `combustible` and `burn` in sense 1. of `ash` and in sense 3. of `coal`.

Lesk reports that his approach has a performance between 50% and 70% on short samples of "Pride and Prejudice" and an Associated Press news story. These numbers, however, are based on "brief experimentations," lacking a thorough evaluation.

4.3.2.2 Yarowsky 1992

A next important step in this line of research was to use a dictionary or a similar resource as a corpus for acquiring statistical models of the most likely contexts of a word (type) by analyzing all dictionary definitions in which the word to be disambiguated appears. Such an approach originates with (Yarowsky, 1992) who used Grolier's Encyclopedia for training of the statistical models. As a set of senses, or rather semantic classes, the categories in Roget's

thesaurus are used. Incidentally, this separates the sense inventory from the corpus used for training, an approach used in most of the subsequent work on WSD.

The basic idea behind Yarowsky's approach is to: 1. collect representative contexts for a particular Roget category; 2. identify salient words within this context; 3. apply the acquired statistical models to predict the appropriate category.

More in detail, consider the following example. In order to disambiguate between two senses of crane (corresponding to the Roget category TOOL or the category ANIMAL), a number of contexts are collected for both these categories. In fact, a context of 100 surrounding words for each occurrence of each member of the category is collected from the Grolier corpus. To avoid a non-representative influence of high frequency words, a weighting scheme is introduced that puts a weight *1/k* on each word, in which *k* is the overall frequency of the word in the corpus. Next, in order to identify salient words in these context concordances, a mutual information like estimate is computed between each context word and the category word in question:

Eq 5: Yarowsky's measure of salience

$$\frac{P(\mathrm{w} \mid \mathrm{Rcat})}{P(w)}$$

The probability of word (w) appearing in the context of a Roget category, divided by its overall probability in the corpus. The log of this salience measure is used also as the word's weight in the statistical model of the category:

Eq 6: Word weight

$$\log \frac{P(\mathrm{w} \mid \mathrm{Rcat})}{P(w)}$$

Some results of training on Grolier`s of the TOOL and ANIMAL categories are listed as follows (salient words, per category, with log of the salience):

animal
species (2.3); family (1.7); bird (2.6); fish (2.4); breed (2.2); ...

tool
tool (3.1); machine (2.7); engine (2.6); blade (3.8); cut (2.6); ...

Selection of the set of words actually used in disambiguation is based on the product of salience and of frequency. “That is to say important words are distinctive *and* frequent.” (Yarowsky, 1992)

Disambiguation is performed by computing a score on occurrences of the selected words in the context of the word to be disambiguated. This score is computed using Bayes’ rule, by summing the weights of these words (*w*) and determining the category (Rcat) for which the sum is greatest.

Eq 7: The disambiguation task

$$\underset{\text{Rcat}}{\text{ARGMAX}} \sum_{w} \log \frac{P(\text{w} \mid \text{Rcat}) \text{ x } P(\text{Rcat})}{P(w)}$$

As an example consider the following window of 10 words around the word `crane`:

(9) `lift water and to grind grain . Treadmills attached to cranes were used to lift heavy objects from Roman times ,`

The table below lists salient words with their individual weights. Summing these, gives clear indication on the `TOOL` sense (or Roget category.) There is hardly any evidence for the `ANIMAL` sense.

Table 1: Salient words with their weights

TOOL	Weight	ANIMAL	Weight
Lift	2.44	water	0.76
Lift	2.44		
Grain	1.68		
Used	1.32		
Heavy	1.28		
Treadmills	1.16		
Attached	0.58		
Grind	0.29		
Water	0.11		
TOTAL	11.30	TOTAL	0.76

On evaluation, Yarowsky reports results on 12 selected ambiguous words that had been previously investigated by other researchers with (some) evaluation results for comparison. Although most of these approaches report between 50% to 84% precision, depending on the level of ambiguity they consider (a distinction between 2 different senses might be easier than that between 5 or more), Yarowsky reports 92% on a mean 3-way sense distinction. Nevertheless, in considering these results, one should not forget the small selection of words on which these results are based. As Yarowsky mentions, “mean performances on a completely random set of words should differ.”

4.3.2.3 Resnik (1997)

As mentioned above, there is strong relationship between WSD and selection restrictions, or *preferences* when seen from a probabilistic point of view. For example, burgundy can be either a color or a beverage. The predicate expressed by the verb to drink, however, will have preference for the beverage interpretation.

In (Resnik, 1997) this link between WSD and selectional preference is exploited in an unsupervised WSD algorithm. The basic idea is to train statistical models of selectional preferences on a POS-tagged and syntactically annotated corpus: the Penn Treebank (Marcus *et al.,* 1993.) Training combines statistical and knowledge based approaches, using WordNet synsets as semantic classes.

The probabilistic model computed in training captures the co-occurrence behavior of predicates and semantic classes in argument position (based on Resnik, 1995.) Taking the example above again, the system computes the selectional preference strength between the predicate drink and the classes beverage and color, defined as follows:

Eq 8: Selectional Preference Strength

$$S_R(p) = D(\Pr(c \mid p) \,\|\, \Pr(c)) = \sum \Pr(c \mid p) \log(\Pr(c \mid p) / \Pr(c))$$

Informally, $S_R(p)$ measures the amount of information that predicate *p* provides about semantic class *c*. This captures the difference between the prior distribution of c and its posterior distribution, given a certain predicate (conditional probability.) For example, in subject position, the prior distribution of class `person` is much higher than that of class `insect`. However, given the predicate `buzz`, the posterior distribution of class `insect` becomes much higher.

Given the definition of selectional preference, a measure of "semantic fit" of a class can be computed that gives an indication of its co-occurrence strength with a particular predicate. This measure, selectional association, can be defined as follows:

Eq 9: Selectional Association

$$A_R(p,c) = 1/S_R(p) \cdot \Pr(c \mid p) \log(\Pr(c \mid p) / \Pr(c))$$

The resulting method for WSD is rather similar to that of Yarowsky (1992), although the approach described here uses linguistic structure as derived from the Penn Treebank and it uses WordNet synsets as semantic classes instead of those found in Roget's. This allows also for an exploitation of the WordNet hierarchy in computing selectional preferences.

As an example, consider two instances of the verb-object relationship in a training corpus, `drink coffee/drink wine`. `Coffee` has two senses in WordNet1.4 and belongs to 13 classes in total (through inheritance in the WordNet class hierarchy.) `Wine` has two senses, too, and belongs to 16 classes in total. Consequently, joint frequencies between `drink` and each of the classes `coffee` belongs to will be incremented with 1/13, and 1/16 for `wine`, which scores will then be used in computing selectional association as described above.[3]

As an example of the predicting and disambiguating power of the approach consider the following selectional association scores for a number of predicates (verbs) and classes (WordNet synsets) in object position:

[3] A somewhat different but effectively similar use of the WordNet hierarchy is described in (Agirre and Rigau 1996). Interestingly, (Peh and Ng 97) tested this algorithm on a domain specific corpus and found it performs under the most-frequent baseline.

Table 2: Selectional Associations (Resnik, 1997)

Verb	Object	Selectional Association	Class
write	letter	7.26	writing
read	article	6.80	writing
warn	driver	4.73	person
hear	story	1.89	communication
remember	reply	1.31	statement
expect	visit	0.59	act

In WSD this information can be used by choosing the class (sense) with the highest selectional association score in case a word has more than one sense. For instance, the word letter has three senses in WordNet1.4: written message (writing), varsity letter, alphabetic character. In disambiguation, the class with the highest selectional association score is chosen.

4.3.2.4 Summary

The unsupervised approaches described above have a number of things in common. They all assume as training data a number of ambiguous words (words belonging to more than one semantic class) with corresponding contexts: Lesk (1986) uses a set of dictionary definitions, Yarowsky (1992) a set of encyclopedia definitions, Resnik (1997) a syntactically annotated corpus of 'general' language sentences. All three of them provide a method of computing, for each of the classes the word belongs to, with which other words and/or semantic classes they are likely to co-occur, given the provided contexts: Lesk (1986) by simply taking the used words in the dictionary definitions, Yarowsky (1992) by computing a relevance for each Roget class (through the words belonging to them) with which the ambiguous word co-occurs in the encyclopedia definitions, Resnik (1997) by computing a relevance, relative to the WordNet hierarchy, for each of the WordNet synsets (through the words belonging to them) with which the ambiguous word co-occurs in the syntactically annotated corpus sentences. Finally, each of the approaches use these computed probabilistic models to predict which of the semantic classes of an ambiguous word will be most likely in a given, novel context, which effectively provides a WSD functionality.

4.3.3 Supervised Methods

4.3.3.1 Semantic Annotation

Supervised methods in WSD assume disambiguated (labeled) data sets on the basis of which probabilistic models can be computed, similar to those used with unsupervised methods. The

significant difference between supervised and unsupervised methods is exactly the availability of such manually annotated training sets, which allows for far more accurate training models.

Manual semantic annotation, however, requires WSD by humans. A number of researchers in fields as diverse as psychology, linguistics and computer science have investigated this task. Jorgenson (1990) found out that human subjects on average are not able to distinguish more than three different senses for a given word. Therefore, a disambiguation task between four or more senses could become problematic. Alshwede (1993) notes that precise identification of a particular sense may not even be possible or necessary. In `The bank closes early on Saturday`, `bank` could refer to both the `building` and the `institution`, which could in fact be listed as two different senses of the word.

Therefore in order to provide a reliable, manually semantically annotated corpus, a number of steps have to be taken (Kilgarriff, 1998): use more than one annotator, calculate inter-annotator agreement (IAA), and determine whether the IAA is high enough. The quest for a high IAA is important, because this forms the upper bound for a WSD program. An automatic program can simply not be expected to perform better on this task than a human. Jorgenson (1990), in the study quoted above, found an average agreement level of about 68%. Fellbaum *et al.* (1997) found that "naive taggers" agreed with experts 75.2% of the time when senses were listed on frequency, they agreed 72.8% of the time when senses were ordered randomly.

The semantically annotated corpus referred to in this research is SEMCOR, which comprises 250,000 words from the Brown corpus and a novel "The Red Badge of Courage" with all content words manually annotated with WordNet senses. Other existing semantically annotated corpora are HECTOR (Atkins, 1993) and the DSO corpus (Ng and Lee, 1996). HECTOR includes 300 word types with between 300 and 1,000 occurrences in a preliminary version of the British National Corpus. DSO covers 192,800 annotated tokens that are the instances of the 191 most frequently occurring and most ambiguous nouns and verbs in a corpus that includes a subset of the Brown corpus and some issues of the Wall Street Journal. All three corpora are English. Semantically annotated corpora for other languages either do not exist or are not publicly available.

In Ng (1999) experiments involving IAA between SEMCOR and DSO are described, which report an agreement of 57% over a subset of 30,315 sentences that are the intersection between the two corpora. In order to investigate the reasons behind this low IAA and to find ways of raising it, additional experiments were carried out to automatically derive coarser sense classes based on the sense tags assigned by two annotators. This process is repeated until an agreeable level of IAA has been reached. Interestingly, this automatic process groups together semantic classes (WordNet synsets) that had been grouped together also completely independently in the context of research into systematic polysemy (Buitelaar, 1998; Tomuro, 2000) -- see also section 1.6.1.1.

4.3.3.2 Methods Used in Supervised WSD

Given a labelled corpus, a number of methods can be readily adopted from the field of machine learning to train a classifying model. In particular, researchers in WSD have used neural networks (Leacock *et al.*, 1993), Naïve-Bayes models (Bruce and Wiebe 1994b, 1999; Gale *et al*. 1992), decision lists (Yarowsky, 1994), and exemplar-based (k-nearest neighbor) algorithms (Ng and Lee, 1996). All of these methods, next to a few others, have been evaluated in a comparative experiment described in Mooney (1996), with Naïve Bayes performing best on the test set used. In the following, (Naïve) Bayesian methods are

discussed in more detail, next to exemplar-based approaches, which have been shown to perform at least as well in more recent experiments (Ng, 1997; Escudero *et al.*, 2000).

4.3.3.2.1 Bayesian Methods

In Bayesian classification, a specified set of *features* is used to make a decision. For WSD, those features are properties of the *context* of the token in question, and the decision to be made is the association of a given token with one of the senses of its type. The performance of the algorithm depends to a large extent on which properties of the context are assumed to be good indicators.

Consider a word *w* with a set of possible senses $S_w = \{s_1, s_2, ...\}$. The general goal is to estimate, for each occurrence of *w* in a context *c*, the conditional probability distribution over S_w, given *c*; that is, the probability $P(s_i \mid c)$ for all $s_i \in S_w$. An application of Bayes' rule to this formula yields

Eq 10: Bayes' rule

$$P(s_i \mid c) = \frac{P(c \mid s_i)P(s_i)}{P(c)}$$

Once this value is obtained for all candidate senses of the token, the sense *s'* is chosen for which it is maximal:

Eq 11: Bayesian decision rule

$$s' = \arg\max_{s_i} P(s_i \mid c)$$

4.3.3.2.2 Modeling the context

What is "the context *c*," and which of its properties should be relied upon in estimating the probabilities used in WSD? The most straightforward view would hold that the context is a sequence of *n* words. In practice, however, there is little hope that this probability can be estimated and put to use without running into sparseness problems. Therefore the notion of context is simplified by introducing *independence* assumptions. At the extreme end of this simplification lies the *naïve Bayes* method, in the NLP literature also known as the *bag of words* approach: Complete independence is assumed between all tokens in the context *c*. That is, both the order in which they occur in the context and the influence that the occurrence of one may exert on the probability of the occurrence of another are ignored.

4.3.3.2.3 Inducing the parameters

The numbers needed in order to apply the Bayesian approach are

1. the prior probabilities of all candidate senses $s_i \in S_w$ for every word w, and
2. for each sense-context pair, the conditional probability of the context, given the sense.

Both of these values can be estimated by Maximum-Likelihood estimation from a labeled corpus.

The right context

So far we have discussed two extreme approaches to the dependencies between contextual features: Assuming that all variables are dependent on each other leads to extremely rich joint distributions and, accordingly, problems with complexity and with data sparseness even in large corpora. The other extreme is the assumption that all variables are independent, which makes the calculations much more tractable, is less liable to suffer from sparseness and over-fitting, but at the same time runs the risk of missing important clues in the decision task. A number of methods have attempted to identify just the right notion of context that maximizes the accuracy on the WSD task while avoiding data sparseness.

Bruce and Wiebe (1994b, 1999)discuss a method that is situated on the spectrum between these extremes. *Decomposable models* of contexts consist of variables that can be grouped into sub-models which consist internally of interdependent variables, but are assumed to be independent of each other. The same utility of independence assumptions motivates the use of Bayesian networks in many other areas of Artificial Intelligence. Bruce and Wiebe (1999) discuss this connection in some detail, borrowing from the graph-theoretical terminology of the literature on Bayesian networks. They discuss a system, which tests all possible independence hypotheses, i.e., all possible configurations of edges in the network, for a given case in an exhaustive search. The goal is to find those models, which describe the data well with just as many interdependencies as necessary.

Furthermore, it is possible to record more than one decomposable model and switch between them depending on certain conditions (Pedersen and Bruce, 1997) or to build context-dependence into the models themselves (Boutilier *et al.*, 1996); the latter has to our knowledge not been implemented in WSD applications. Bruce *et al.* (1996) and Bruce and Wiebe (1999) provide a good discussion of the ways in which feature selection, model properties, and probability estimates may be isolated and tested independently.

In an experiment on 34 words of the HECTOR corpus (Atkins, 1993; Hanks, 1996) chosen in accordance with the design of the SENSEVAL competition, the Model Selection algorithm described by Bruce and Wiebe (1999) generally performed as well or better than the naïve Bayes method. The accuracy rose to an average of 76.4% from 75.0% for naïve Bayes. It was significantly better ($p<0.05$) for six of the words and significantly worse for none of them. Moreover, by selecting the best models considered during the process by the algorithm, the accuracy can be brought to an average of 80.8%, a fact which shows that the model selection and evaluation process seems to have potential for further improvement. Bruce and Wiebe also discuss differences between the words on which the model selection algorithm performed significantly better and those on which it performed worse, observing that the former have a higher average number of candidate senses and are “harder” (have a higher average entropy.)

This shows that it is the highly frequent and contextually versatile words for which it is most fruitful not to ignore the most salient dependencies between features. On the other hand, the considerable computational overhead, at least at the current stage of technological development, may in practical applications offset the rather modest increases in accuracy afforded by this method.

Other contextual features

The foregoing discussion considered only approaches in which the context is defined by the size of a text window surrounding the target token. These methods have been refined, for example, by incorporating other kinds of information, such as the part of speech of each item in the context, its grammatical relation to the target token, or its distance from it (cf. the distinction in Leacock *et al.* (1998) between "local" and "topical" context.) The TLC system presented by Chodorow *et al.* (2000) can be configured to distinguish between open-class and closed-class words and treat them differently depending to their position. Other work investigating such mixed approaches includes (Hearst, 1991; Yarowsky, 1993; Bruce and Wiebe, 1994b; Pedersen and Bruce, 1997; Leacock *et al.*, 1996; Leacock *et al.*, 1998).

4.3.3.3 Exemplar Based (Instance Based, k-Nearest Neighbour)

Exemplar based methods center around the similarity between two instances. Given a labelled corpus, all training examples are kept in memory and in classification the test instance is compared to all of these. Then, the *k* training examples with the shortest distance to the test instance are determined and the class with the majority among these will be assigned as the class to the test instance.

Ng and Lee (1996) use an implementation of this basic algorithm for two WSD experiments, one focusing on the word `interest`, using data collected by and comparing results with Bruce and Wiebe (1994), and one on a set of 191 nouns and verbs in the DSO corpus which had been constructed specifically for the experiment.

For the first experiment, the average accuracy over 100 random trials was at 87.4% much better than the accuracy of 78% reported by Bruce and Wiebe (1994). Results of the second, larger experiment were significantly higher than the "Sense 1" and "Most Frequent" baselines, but somewhat lower than Naïve Bayes. (The first baseline can be obtained from the ordering of senses in WordNet, the second from the labelled instances in the semantically annotated training corpus.)

However, in further experiments, better results were obtained by increasing *k*, the number of most similar instances used in classification. With k=20, WSD results increase about 4 points compared to k=1. Additionally, by use of 10-fold cross validation, *k* was set automatically to a certain optimum, which increased results slightly further, beating Naïve Bayes by a slight margin.

In order to further investigate the reasons behind a possible superiority of exemplar based over Naïve Bayes, Escudero *et al.* (2000) performed an additional number of experiments that include several settings for *k* and example and attribute weighting. The results of these experiments showed that exemplar based will be at least as good as Naïve Bayes. If an

additional metric (Modified Value Difference Metric) is used that allows graded guesses of the match between two different symbolic values, then exemplar based will be even better.

Importantly, however, these results seem to depend on the nature of the attributes. If a large context window was used, a sparse data problem occurred which did not influence Naïve Bayes, but degraded exemplar based significantly. But this observation should perhaps not be surprising, given the independence of attributes with Naïve Bayes and the implicit dependence of attributes with exemplar based instances on the other.

4.3.3.4 Semi-supervised Methods

In supervised methods such as those discussed above, the need for prepared training corpora large enough to avoid problems with data sparseness still constitutes a *bottleneck*, even though the rules for the decision procedure can be inferred automatically by statistical observations. There are various ways in which the bottleneck at the training stage has been addressed. The most radical of these is completely unsupervised *clustering* of the occurrences in an unlabeled corpus. We will discuss this in the next section on purely empirical approaches.

Short of working without any tagging of the training data, one can use a small amount of sense data gleaned from a knowledge source to *seed* an algorithm which then extends the categorization to all occurrences in the training set. Karov and Edelman (1998) describe one such approach. A combination of two dictionaries (Webster's and the Oxford dictionaries) and a thesaurus (WordNet) was used to seed the algorithm (see also Yarowsky (1995) for a similar approach).

The training corpus is augmented with "feedback sets," consisting, for each word *w* and sense s_i in S_w, of those words of the definition of s_i which do not occur in the definitions of other senses of the *w*. Each sentence *s* containing a target word *w* is called an *example*, or *context*, of *w* and is assigned a similar set of features, consisting of the words in *s* and those occurring in the definitions of *w*.

Two similarity measures are defined and refined in a stepwise fashion: Similarity between words is proportional to the similarity of the sentence in which they occur, and similarity between sentences in turn is defined by the similarity between the words which occur in them. An iterative algorithm starts out by treating words as similar only to themselves and subsequently refining this measure, relying on an auxiliary notion of *affinity* between words and sentences and a complex term weighting scheme. During each iteration, each feedback set can be augmented with new sentences "attracted" by (i.e., similar to) one of its members. When the iterative algorithm has converged, the resulting sets of sentences for each sense are partitioned into *typical use sets*, grouping together those sentences, which were "attracted" by the same sentence during the iterations. In disambiguation, a sentence *s* containing the target word is assigned the sense containing the "typical use" cluster most similar to it.

Karov and Edelman tested their algorithm on 500 occurrences of each of the words `drug`, `sentence`, `suit`, and `player`, using the Treebank-2 corpus (one million words) and the dictionaries mentioned above for training. The average correctness on the four words was roughly 92%. No baseline was given.

4.4 Empirical Approaches

As we said above, the two kinds of knowledge involved in WSD are

1. the set of candidate senses for a given type, and

2. the contextual clues used in deciding which of its senses a word assumes in a particular occurrence.

The methods discussed so far presuppose that at least one of these kinds of information is provided by an external source. As we saw this source may be a structured semantic database, such as a dictionary or thesaurus, or a hand-tagged corpus from which the information needed in (2) is obtained.

In the previous section we discussed statistical methods which use such tagged corpora and subsequently apply the results to the *decision* problem of choosing between the possible senses of an occurrence. Although the use of statistics for those purposes does lead to high accuracy on the decision task, the fundamental problem of the *bottleneck* still arises from the the need for large amounts of prepared training data.

Moreover, systems relying on such externally provided knowledge have limited *portability*: Domain-specific requirements for finer or coarser grain in the sense distinctions and definitions can only be met to the extent that the tagging of the training data relects those requirements. For instance, in an application built specifically for the medical domain it may be necessary to make fine sub-distinctions between the names of diseases. Thus for a word like `tumor`, it may be useful to exploit the distinction between the `benign` and the `malignant` readings as a real ambiguity. General-purpose knowledge sources, however, may well fail to list them as separate senses. On the other hand, some commonly made distinctions may be irrelevant: It would be of little value in medical information management to give the sense of `bank` used in games (as in `to hold the bank`) the same ontological prominence as that of `blood bank`. In many general dictionaries, however, the first is given a separate definition, while the second is not.

These difficulties motivate the application of unsupervised methods not only in classification, but also in *discovering* the candidate senses in the first place. The basic assumption is, as before, that the context surrounding a word occurrence is highly indicative of its sense. The task is now to examine the contexts of all occurrences of the word in question. If those contexts fall into multiple clearly distinguishable classes, the word is treated as ambiguous and the context classes can be used as representations of its senses.

Thus the goal is to discover patterns in the collocational behavior of words. From a Machine Learning point of view, the task is not classification, but *clustering*.[4] As such, it requires two main ingredients:

1. a measure of association, or similarity, between contexts, and

2. an algorithm using the measure to group similar contexts together.

[4] General discussions of clustering methods in NLP and IR can be found in Duda and Hart (1973), Rasmussen (1992), and Manning und Schütze (1999).

The measure of similarity between contexts is often based on, and related to, a measure of similarity between words. It is therefore useful to start with a discussion of some of the approaches to the latter, to the extent that they have been applied to WSD. The most prominent of them is what we will call the *vector space* model.

4.4.1 Vector-based approaches

In the vector space approach to word similarity, the collocational properties of words are represented in a high-dimensional space. Similar vector representations are the most commonly used technique for assessing the similarity between documents and between queries and documents in Information Retrieval (IR; cf. the chapter on CLIR.) In this section we provide a basic outline of the approach. A discussion of some details is provided below.

For word similarity, a vector is associated with each word type *w* which represents the co-occurrence pattern between *w* and a selected set of indicative, content-bearing words. Those content-bearing words label the *dimensions* of the vector, or, the *features* whose values are relevant for the characterization of the word. The full vector space is represented by the matrix $V \times C$ for a vocabulary *V* of lexical items and a set *C* of dimension labels.

Below we will discuss different ways of filling in the cell of the matrix. First, however, to give a general idea of the way such a matrix is utilized, we assume that it is given

4.4.1.1 Word similarity

Similarity between two vectors in a matrix as described above is measured as the *normalized correlation coefficient,* which in geometric terms is equivalent to the *cosine* between the vectors. Suppose two words *w, v* are represented by two *n*-dimensional vectors $\vec{v}, \vec{w}$. Then the correlation is calculated as in Eq 12: Vector similarity.

Eq 12: Vector similarity

$$\mathrm{corr}(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2 \sum_{i=1}^{n} w_i^2}}$$

This value lies between 0 and 1; the higher it is, the closer to each other the vectors are situated in the space and, by assumption, the more similar the words are semantically.

4.4.1.2 Context vectors

The goal is, however, to measure the similarity, not between words, but between the *contexts* of word occurrences. To this end, context vectors are calculated from the vectors of the

words constituting the contexts by summing up the vectors, optionally giving them a *weight* to account for their informativeness.

Context vectors are thus of the same type as word vectors, and similarity among contexts (as well as between words and contexts) can again be measured as the cosine between vectors, defined in Eq 12.

4.4.1.3 Sense vectors

In vector based approaches, the set of senses each word type can assume are represented as vectors in the same space as words and documents. If the method is partly supervised, that is, if a labeled training corpus is available, then the contexts of all occurrences of each sense can be stored in a list or added up to a vector representing that sense. Niwa and Nitta (1994) prefer the former over the latter, arguing that "a pecularly similar example is more important than the average similarity" (p. 306).

In fully unsupervised methods, no labeled training data are available, and the division of senses for each type must itself be induced using statistical methods (clustering of the word vectors) based on unlabeled text. This is the approach of Schütze (1997, 1998,) which we will discuss in more detail below.

4.4.1.4 Disambiguation

Once the sense vectors have been obtained for each ambiguous word *w*, the actual disambiguation is again a classification problem as described in the section on (partly) supervised methods. The sense vectors represent the candidate senses s_1, s_2, ... of word type *w*. For each occurrence *t* of *w*, the algorithm

1. maps *t* into its context vector $\vec{c}_t$
2. retrieves the set s_1, s_2, ... of sense vectors for *w*
3. chooses among the sense vectos the one which is closest to $\vec{c}_t$ and assigns it to *t*.

4.4.2 Implementations

The preceding sections give an overview of the use of vector-space approaches in WSD. We will now summarize some of the work that has been done in this area. There are a number of ways in which the basic notions have been fleshed out.

4.4.2.1 Choosing the column labels

One parameter influencing the performance of a vector based similarity model is the choice of the set of "content-bearing" words used as dimension labels.

4.4.2.1.1 Global selection

Niwa and Nitta (1994) report an experiment comparing a partly supervised method based on a dictionary with a fully unsupervised, vector-based representation of word similarity. As dimensions were chosen the 51st through 1050th most frequent word in the Collins English Dictionary (CED). The vocabulary of the corpus they used (the 1987 Wall Street Journal) comprised about 31,000 types.

A simple frequency cutoff may also be used to choose a smaller set of content words. For instance, Flournoy *et al.* (1998), in an IR experiment, eliminate the members of a stoplist and then choose the 50th through 1049th most frequent words of the remaining vocabulary. In one of the experiments described by Schütze (1998), the 2,000 most frequent non-stoplist words are chosen. to build a matrix of 20,000 vectors with 2,000 dimensions each.

These selection methods are what Schütze called *global*: Dimension labels are chosen based on properties of the whole corpus, and the same resulting space is used to build context representations for all word types.

4.4.2.1.2 Local Selection

Schütze also implemented a *local* selection criterion: For each word *w* that is to be disambiguated, only that subset of the whole vocabulary is considered whose members actually cooccur with *w*. As a consequence, the set of dimension labels, as well as the row labels in the matrix (i.e., the set of words for which vectors are constructed,) vary depending on *w*. The effect of this is that the corpus is reduced to just the contexts of *w*, ignoring everything else. In Schütze's experiment, a symmeteric matrix of 1,000 by 1,000 dimensions was built in the local selection method.

Schütze's experiment with local selection also uses a criterion different from the simple frequency cutoff. Instead, a χ^2-measure is employed which is high for words who tend to occur mostly in the vicinity of the word in question. It is defined in Eq 13: Given a word *w* to be disambiguated, for each word *v* which occurs in the contexts of *w*, where n_{++} is the number of contexts in which both *w* and *v* occur, n_{--} the number of contexts in which neither *w* nor *v* occur, and so on.

Eq 13: χ^2-measure for local feature selection

$$\chi^2 = \frac{n(n_{++}n_{--} - n_{+-}n_{-+})^2}{(n_{++} + n_{+-})(n_{-+} + n_{--})(n_{++} + n_{-+})(n_{+-} + n_{--})}$$

In the outcome of Schütze's experiment, global selection was superior to local selection (see below.)

4.4.2.2 Filling the cells

Niwa and Nitta (1994) built a vector for each word *w* in the vocabulary whose 1,000 cells c_i were filled with an estimate of *mutual information* between *w* and column label *i*. In disambiguation, the set of candidate senses for each target word is assumed given, and only two senses per ambiguous words are tested in the evaluation.

A more commonly employed method is to fill the cells in the matrix directly with co-occurrence counts. The first question to arise then is how co-occurrence should be defined. This problem is similar to that of defining what a context is in supervised training, which we discussed above. Currently, fixed-width rectangular text windows are most commonly used. Schütze (1997) found that with his model, a fixed-width window of ±50 tokens yielded the best performance on WSD. Burgess and Lund (1997) use a window of 10 words.

Other possibilities would be to use a flexible window spanning a certain number sentences, paragraphs, the document, or a thematically coherent text segment. Automatic methods for determining the latter have been explored in a number of ways (Kozima, 1994; Hearst, 1997; Reynar, 1998; Kaufmann, 2000) and have been evaluated in the Topic Detection and Tracking (TDT) project (TDT 1997,) but have not been implemented in WSD systems.

Burgess and Lund (1997) produce a quadratic matrix representing the full cross-product of the 70,000-word vocabulary with itself, that is, each item in that vocabulary serves both as a row label and a column label in the matrix. They note, however, that far fewer dimensions (as few as 1,000) serve almost as well for their purposes.

4.4.2.3 Dimensionality reduction

Another IR technique whose application to word vectors has been shown to be fruitful is Singular Value Decomposition (SVD; cf. Berry, 1992; Golub and van Loan, 1989,) a dimensionality reduction technique mapping the high-dimensional vector space to one of lower dimensionality in which the relative similarities between word vectors are preserved as well as possible. In Information Retrieval, SVD is applied in Latent Semantic Indexing (LSI; cf. Deerwester *et al.*, 1990.) Aside from the gain in processing speed, the benefit of applying SVD to vectors of co-occurrence counts is twofold: The resulting real-valued matrix is less sparse, and the similarities and dissimilarities between word vectors are amplified. For details, see the discussion in Schütze (1998).

4.4.3 Clustering

Given the set of all context vectors for a given ambiguous word, the role of clustering lies in discovering patterns in their distribution in the vector space. The assumption is that the separate senses will occur in thematically separate contexts, so that the vectors of those contexts will be located in distinct areas in the space. There is much previous work on clustering, but little in the area of inducing word senses from context grouping. Schütze (1997, 1998) uses the *Buckshot* algorithm (Cutting *et al.*, 1992), a combination of the Expectation-Maximization (EM) algorithm and Group-average Agglomerative Clustering (GAAC). The latter is used on a sample from the vocabulary to seed the EM algorithm, which would otherwise be liable to converge on a local maximum. For a formal

characterization of the application, see Appendices B and C of Schütze (1998) and the references therein; here we describe only the general idea.

As mentioned above, the measure of similarity between clusters is the same as that between words, *viz.* the cosine, or normalized correlation coefficient, between the corresponding vectors. A "good" cluster Γ is one whose members have a high average correlation $C(\Gamma)$, as defined in Eq 14.

Eq 14: Average correlation

$$C(\Gamma) = \frac{1}{2}\frac{1}{|\Gamma|(|\Gamma|-1)}\sum_{\vec{v}\in\Gamma}\sum_{\vec{w}\in\Gamma}\mathrm{corr}(\vec{v},\vec{w})$$

The algoritm proceeds by *merging* clusters iteratively, at each step keeping the larger cluster with the highest average correlation. Initially each context vector is treated as a separate cluster. The algorithm stops when a pre-determined number of clusters. In Schütze (1998), two experiments are reported with two and ten clusters, respectively. Evaluation was performed only on two-way ambiguous words, however.

4.4.4 Results

4.4.4.1 Niwa and Nitta (1994)

Recall that Niwa and Nitta built word vectors based on co-occurrence counts (similarly to Schütze's, but without dimensionality reduction) and compared that method to the "link length," or distance-based approach which uses dictionaries as knowledge bases. As we said, the disambiguation step in their experiment cannot be properly called unsupervised, as they used a hand-picked set of sample contexts for training (20 per sense.) These two methods were evaluated on nine words with two senses each, counting as "context" the vocabulary in text windows of size up to ±50 words. The results were only presented graphically, with no exact numbers of precision. Precision seems to range form around 70% (for order) to nearly 100% (for suit.) As Niwa and Nitta point out, however, it is obvious from the graphs that the method using co-occurrence vectors is superior to the dictionary-based distance measure.

4.4.4.2 Schütze (1998)

Schütze's experiments with 10 "pseudowords" (cf. Section 4.6.3 below) and 10 actual ambiguous words explored a number of the options mentioned in the preceding section. Table 3 reproduces some of the results. The heading labels are meant as follows:

1. Local *vs.* global selection of dimension labels

2. χ^2 values (cf. Eq 13) *vs.* frequency counts in the vector cells
3. Term vectors *vs.* SVD-reduced vectors
4. Two *vs.* ten clusters.

The meanings of these distinctions are explained above.

Table 3 : Results of Schütze's experiments

Local								Global	
χ^2				Frequency				Fequency	
Terms		SVD		Terms		SVD		SVD	
2	10	2	10	2	10	2	10	2	10
72.1	77.9	84.1	88.5	77.8	81.8	82.9	88.3	89.7	90.6

The table shows that the rightmost column yielded the best accuracy. This suggests that global feature selection, frequency counts as vector-cell values, and SVD reduction yield the best design for a system like this. For more statistical information and discussion, see Schütze (1998**)**.

4.5 Cross-Lingual Matters

From a cross-lingual point of view, translations of a word in another language can be taken as indicative of different meanings of this word. This observation has been exploited by a number of researchers in WSD to avoid the unavailability of labelled corpora. Instead they took advantage of the existence of readily available parallel corpora in two or more languages. For instance, Gale *et al.* (1992b) discuss a system based essentially on Naïve Bayes, using the Hansard Corpus, a parallel corpus of Canadian parliament reports in English and French, in which they had aligned sentences automatically. They then used the French translations of the six English test words as training data, on which the algorithm attained an accuracy of roughly 90%: `duty`, `drug`, `land`, `language`, `position`, and `sentence`. These six words were chosen because they happen to have senses which are translated into different words in French.

Table 4: English words disambiguated by their French translations (Gale *et al.*, 1992b)

ENGLISH	FRENCH	SENSE	ENGLISH	FRENCH	SENSE
duty	droit	tax	language	langue	medium
	devoir	obligation		langage	style
drug	médicament	medical	position	position	place
	drogue	illicit		poste	job
land	terre	property	sentence	peine	judicial
	pays	country		phrase	grammatical

Obviously, research in this area is closely related to corpus based approaches in cross-lingual information retrieval as discussed in Chapter 2. In this area of research, multilingual dictionaries are generated automatically from parallel corpora and then used in information retrieval to translate query terms and/or key phrases in the documents to be retrieved.

Interestingly, however, parallel corpora can also be used to tune existing lexical semantic resources used for WSD such as WordNet (Resnik and Yarowsky, 1997; Ide, 1999). The experiment described in (Ide, 1999) is on a small multilingual parallel corpus: George Orwell's "1984" in English, Slovene, Estonian, Romanian and Czech. This text consists of about 100.000 words and has been translated out of English directly in each of the other languages. For the experiment, nine ambiguous English words were considered: `hard`, `head`, `country`, `line`, `promise`, `slight`, `seize`, `scrap`, and `float`. The occurrences of these words were annotated with WordNet senses for English and then provided with alignments for the other languages by native speakers. The experiment was then to compute a so called "Coherence Index" (CI) between different WordNet senses, in order to see how consistently they were translated with the same words in the other language(s). A CI of 0 indicates that two senses are translated by different words each time, whereas a CI of 1 indicates that the same word is used consistently for both senses. The CIs computed between each of the senses of a word are then used to cluster these. Two senses will be clustered if their CIs are high. This produces a hierarchical clustering of senses that mirrors the division of senses in some dictionaries. Such a hierarchical sense structuring is, however, not available in WordNet and this method could therefore be used to provide this in an empirical way (and therefore adjustable to certain domains and/or applications).

From the point of view of WSD this may be important information, because it produces an indication of how closely related two senses are and therefore how likely they are to be (automatically) distinguishable. Unrelated senses will be more easily handled by an automatic WSD system. (See also the next section on evaluation.)

4.6 Evaluation

While the practical and theoretical work in the area of WSD carried out during the nineties produced and refined an impressive variety of methods, it was, and to a large extent still is, difficult to compare the published results against each other reliably. Many focus on a handful of selected words, which either happened to strike the interest of the author or were discussed in the previous literature. This proliferation of arbitrary test items throughout the literature has led to informal "standard" cases, such as the words `line`, `serve`, and `hard`, which

figure prominently in some articles of the March 1998 special issue of *Computational Linguistics* on WSD.

Even if performance on the same words is compared between authors, however, there are still many degrees of freedom undermining the validity of the comparison. Rarely are exactly the same training and test data used. The tag sets and even the "right answer" may vary considerably from author to author. Furthermore, if the corpus is pre-processed in any way that is prone to introduce errors (such as part-of-speech tagging, shallow parsing, or manual tagging,) it is impossible to tell whether differences in performance are solely due to differences in the WSD part of the systems.

4.6.1 Measures of accuracy

A measure of the extent to which two or more assignments of labels to tokens in text agree is indispensable both in order to evaluate the quality of a "gold standard" distilled from the judgments of several human annotators, and to measure the performance of systems against such a gold standard. One wants to obtain a normalized measure *A*, either expressed as a real number between 0 and 1 (inclusive) or in terms of percentages.

4.6.1.1 Exact-match measures

The most simple-minded such measure would be to count matches against mismatches:

Eq 15: "Exact match" criterion

$$A = \frac{\#(\text{correctly assigned labels})}{\#(\text{assigned labels})}$$

There are two problems with this measure. The first lies in the fact that depending on the number of sense labels and words to be disambiguated, the baseline, in this case the agreement that would be expected to arise by chance if the assignments were made at random, may differ. In labelling tasks in general, a now commonly employed method of controlling for this factor is the *Kappa* measure (Carletta 1996): Let *P(A)* be the actual agreement and *P(E)* the expected agreement by chance, obtained by considering all occurrences of labelled words random variables over which tags are randomly distributed. Then

Eq 16: The Kappa measure

$$K = \frac{P(A) - P(E)}{1 - P(E)} = \frac{\text{unexpected agreement}}{\text{expected disagreement}}$$

In other words, if, for instance, the chance agreement would be expected to be .5, then taking the *Kappa* measure amounts to re-distributing the probability mass for the analogue of Eq 15 to the region above .5.

The second problem with this approach runs deeper, however, and requires a more fundamental rethinking of the task itself. Both Eq 15 and Eq 16 are applicable only in cases where exactly one candidate assignment is "right" and all the others are "wrong." In WSD, this view would correspond to the "checklist" theories of Fillmore (1975). There is widespread agreement that that view is misleading, for obvious and not-so-obvious reasons. It is obvious that many sense labels form hierarchies: The financial sense of `bank` is in a way an underspecified super-sense of both the `building` and the `corporation` senses, which are obviously related: a corporation has its seat in a building. Identifying such systematic relations between senses, leading to sense hierarchies, has been the topic of recent research in computational lexical semantics, building on WordNet and other lexical semantic resources (Dolan, 1994; Buitelaar, 1998; Peters *et al.*, 1998; Tomuro, 2000). So, if a system chooses a more general, underspecified sense where one of the more specific ones is given in the "gold standard," it should not be penalized with the same severity as for other mistakes.

The less obvious sense in which the "checklist" theories are misleading is due to the fact that even at the same level in the hierarchy, treating senses as mutually exclusive is too crude an approach. What seems more appropriate is to think of the "sense potential" of a word as comprising a number of features which may or may not be invoked, to differing degrees, in a given a occurrence. An illuminating recent discussion of this can be found in Hanks (2000). The consequence of such a view, were it to be upheld in practice, would be that senses are considered inherently vague, "match" in WSD is *always* a matter of degree, and the very practice of giving bipolar solutions in the gold standard is on the wrong track. To be sure, it seems that it would be much more difficult to reach agreement on the gold standard if its prescriptions were gradient, rather than discrete.

But things have not developed so far. The SENSEVAL-1 competition, to be discussed below, does not employ a vague notion of senses in the gold standard. Its evaluation metric, however, does address the "checklist" problem.

The problem of imposing the all-or-nothing matching approach on current WSD systems is especially apparent given that many of those systems work internally with gradient measures of probability, similarity or the like, and in most cases will not assign a zero score to the correct answer, even if they end up deciding on an incorrect one. In practice, embedded in applications, such systems are rarely forced to make a decision at the WSD stage. Rather, their output is used as-is to be later combined with information from different sources. An evaluation which does not give some credit for the non-zero probability such a systems does assign to the correct answer would draw a distorted picture of its usefulness.

4.6.1.2 Cross-entropy

Resnik and Yarowsky (1997, 1999) suggest that the best measure in such circumstance would evaluate the *probability distribution* obtained by the WSD system without forcing it to make a decision. This would require a measure of the difference between the "distribution" over the candidate senses in the gold standard (which in SENSEVAL-1 assigns 1 to the "correct" one and 0 to all others) and the distribution obtained by the machine. A *cross-entropy* based measure suitable for this is obtained as in Eq 17. Here N is the overall number of tokens disambiguated, and $P(s_i|w_i,c_i)$ is the probability assigned by the system to the correct sense s_i of word w_i in context c_i.

Eq 17: Cross-entropy based score

$$A = -\frac{1}{N}\sum_{i=1}^{N}\log_2 P(s_i \mid w_i, c_i)$$

This measure, which may be employed with or without the log (Resnik and Yarowsky, 1999), rewards systems which assign a relatively high probability to the correct senses, even if they do not select it in all cases.

4.6.1.3 Semantic distance

Another problem addressed by Resnik and Yarowsky (1999) is that even if a sense assignment is clearly "off the mark," it may be more or less so. Returning to the hierarchical organization of many areas in the sense tag space, the penalty for cases in which two close siblings in the hierarchy of senses are confused could be alleviated compared to cases in which a grossly different sense was chosen. This presupposes a measure of *distance* between word senses, which Resnik and Yarowsky propose to calculate either from the hierarchical organization of one or more dictionaries or from the lexicalization of the senses across languages. Inherently similarity-based sense representations, such as the vector space model, are not discussed in the article.

Given such a measure of distance, one could simply strive to minimize the average aberration over all cases. But a more sensible use of the measure might be as a weight in Eq 17. The idea is to weigh misappropriated probability mass the heavier, the "farther away" it is from the correct assignment according to the similarity measure. In Resnik and Yarowsky (1999), this is given (without the log) as in Eq 18, where s_i is again the correct sense and s_j ranges over all candidate senses for word w_i.

Eq 18: Distance weighting

$$A = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in S_{w_i}}\mathrm{dist}(s_i, s_j)P(s_j \mid w_i, c_i)$$

The penalty for "far-out" probabilities will under this measure be amplified.

4.6.1.4 SENSEVAL-1 conventions

Melamed and Resnik (2000) added a number of extensions and applications of these ideas, which were adopted in the SENSEVAL-1 competition. They concern the use of non-probabilistic data in calculating probabilistic scores and were motivated in part by practical problems in SENSEVAL. We will only briefly mention them in this section.

The formulas in the previous section are applicable in cases in which WSD systems output not determinate choices of sense labels, but probability distributions over the candidate set. The case in which exactly one sense is chosen is a special case of this which can be modeled with an irregular distribution, assigning 1 to the choice and 0 to all other candidates. Some systems, however, put out *sets* of senses. In those cases, the probability mass was distributed uniformly over the members of the output sets.

On the other side, the gold standard in some cases provided more than one "correct" tag. Those sets were interpreted disjunctively, assuming uniformity in the "correct" distribution. That is, the probability mass assigned to *any* of the candidates contributed equally to the score:

Eq 19: Multiple correct tags in SENSEVAL

$$A = \sum_{s_j \in S_i} P(s_j \mid w_i, c_i)$$

In Eq 19, S_i denotes the set of correct labels given for the occurrence of word w_i.

Finally, an approach to introducing a measure of distance based on the hierarchical organization of the tag set was introduced, based on the idea that an assignment of a label which stands in a parent-child relationship to the correct answer is "partially right." The way this was addressed in SENSEVAL is by assuming a uniform distribution over the child nodes at every branching point in the hierarchy. If node p has children $c_1,...,c_n$, then the "upward" and "downward" probabilities are as given in Eq 20, for $0 \le i \le n$.

Eq 20: Probabilities in hierarchical tag sets

$$P(p \mid c_i) = 1$$

$$P(c_i \mid p) = \frac{1}{n}$$

In the evaluation, all probability was then distributed to the leafs of the hierarchy, using Eq 20, in both the gold standard and the output of the WSD systems, as well as in the calculation of inter-rater agreement using the *Kappa* statistic.

4.6.2 Senseval

The SENSEVAL competition, held for the first time in summer 1998, is conceived as a way of rallying the researchers in the field around a common view of the fundamental problems to be addressed, while yielding reliable data as to what works and what does not, generally accepted evaluation metrics and benchmarks. A special issue of *Computers and the Humanities* (vol. 34, no. 1/2, April 2000; henceforth CH) is devoted to the design, results, and general considerations around that first tournament.

4.6.2.1 General issues

Kilgariff and Palmer (2000) take pains to discuss objections which were or might have been encountered along the way when such a standard was proposed. Some of those objections are common to similar competitions and concern social factors such as the attitudes towards risk-taking and collaboration, and the notion of progress fostered by the project. Others address more fundamental objections. The project can be divided largely into two sub-tasks:

1. Defining the task, and

2. Producing a "gold standard."

The second point depends to a large extent on the answer to the first. (1) raises a number of controversial issues on which general agreement has traditionally been hard to attain.

The major technological question is one of modularity: whether it is sensible, or even possible, to isolate WSD as an NLP task in its own right and, if it is agreed that it is, how to define it. Here it is important to keep in mind that in the greater landscape of NLP *applications*, WSD has no place, but figures merely as an enabling task. An NLP system into which a WSD component is organically integrated may perform well while the WSD component in isolation may be of little use.

But there is also the more fundamental theoretical problem that the utility of a definition of a subfield depends on the validity of the theory on which the definition is based. It is fair to say that we are far from having a generally acceptable theory of word meanings. Lacking that, how can we be sure that SENSEVAL addresses the right problems in the right way?

Fortunately the organizers are themselves keenly aware of these pitfalls; cf. the discussions in Kilgariff and Palmer (2000), Hanks (2000) and elsewhere in the special issue of CH. Therefore there is hope that SENSEVAL will be allowed to evolve with, rather than preempt, theoretical developments, and thereby secure its role as a major forum for evaluation and comparison in the field.

4.6.2.2 SENSEVAL-1 (1998)

The first SENSEVAL competition was held in 1998. It covered English, French and Italian; the latter were collectively treated under the heading "ROMANSENSEVAL." Here we only briefly summarize the English component. A complete description of the event and some short papers by its contributors were published in the April 2000 issue of the journal *Computers and the Humanities* (volume24, no. 1/2).

The English SENSEVAL-1 had 18 participants. Both the corpus and the dictionary used were taken from the outcome of the HECTOR project (Atkins, 1993), a dictionary cross-linked with text samples; the latter were used as training and test corpus. The task was to disambiguate a sample of 40 words, which were obtained by randomly choosing 10 words from each of four frequency classes. The gold standard was developed by professional lexicographers, aiming at a high level of inter-rater agreement. Agreement was at an average of 95.5% according to the Kappa measure (cf.

Eq 16 above.) This high inter-rater result may be thought of as an upper bound on the performance of the participating computational systems.

Scoring was performed at three different levels of "granularity," using the scheme described in Section 4.6.1.4 above. At the fine level, only exact matches were counted as correct. At the other extreme, the hierarchical tag set was simplified by treating everything below high-level senses as equivalent, collapsing fine distinctions. At the intermediate level, the probability mass of a parent node was distributed uniformly over its children, as described above. As noted by Kilgariff and Rosenzweig (2000), the distinction made little difference in the overall ranking of the systems.

The results of SENSEVAL will not be reproduced here, referring the reader to the discussion and graphs in Kilgariff and Rosenzweig (2000) instead. Overall, the accuracy of the best systems was around 74-78%.

4.6.3 Unsupervised evaluation: Pseudowords

From the discussions so far it is clear that evaluation involves its own *bottleneck* of knowledge acquisition, due to the cost and effort required to produce reliable hand-tagged training and test data. It is hoped that the SENSEVAL competitions will over time generate large amounts of such data sets and make them available.[5]

In the absence of such high-quality rescources, a simple method to produce data sets from unlabeled corpora is to generate "artificial" ambiguity, as it were, by collapsing semantically clearly different words into unique strings. The method has been discussed by Gale *et al.* (1992c) and Schütze (1992, 1997). Schütze, for example, created the nonsense word `banana_door`, made a copy of the corpus in which every occurrence of `banana` and `door` was replaced with the new string, and ran his algorithm on the new corpus with the goal of recovering the original words. Schütze (1998) uses pairs of words instead of single words to create strings out of items like `wide range` and `consulting firm`.

[5] The resources of SENSEVAL-1 were put in the public domain in August, 2000 at http://www.itri.bton.ac.uk/events/senseval/ARCHIVE/resources.html.

Pseudowords are particularly useful for algorithms requiring large amounts of data, such as the unsupervised clustering method discussed in Section 4.4.3 above. On the other hand, the fact remains that the ambiguity thus created is *artificial*. There is no simple way to mimick truly ambiguous words, whose senses may differ to various degrees and be ordered hierarchically, by collapsing words in this way. Therefore we see a good possibility that the use of pseudowords will be abandoned with the availability of more reliable large data sets.

5 The Medical Domain

5.1 Overview: Basic Principles

5.1.1 History of Clinical Classification and Terminology

The implicit abstractions in a concept are, according to Plato's articulation of a perfect form, apart from the shadow of that form in this world. Plato classified things by using strict divisions, a method rejected by Aristotle. But abstract descriptions are not words or terms. Nomenclatures, on the other hand, conventionally can be nothing more than lists of recognized or sanctioned words and have little or no relationship with a system of classification. Many authors invoke terminology to subsume language labels for concepts of the entire problem, from classifications to nomenclatures.

Charles S. Peirce (1839 - 1914), an American logician, defined the semantic or semiotic as: "an action, an influence, which is, or involves, a cooperation of three subjects, such as sign, its object, and its interpretant, the relative influence of the three not being in any way resolvable into actions between pairs"(semiotic triangle.)

The first person who developed a patient history was Hippokrates (Greek medical doctor, ca. 460 a. c.-370 a.c.), the founder of scientific medicine.

Most terminologists credit the birth of the Standardizes Nomenclature of Diseases (SND), which was later to become the SNDO (with Operations), as the beginning of a modern era for clinical description. SNDO introduced in 1929 as a multi-axial coding system. The two axes were topology (or anatomy) and etiology (or pathophysiology.) An evolutionary ordering would be SND, SNDO, SNOP (pathology), SNOMED (medicine), SNOMED II, SNOMED International, SNOMED RT, and the recently announced product of the merge with the U.K. National Health Service's Clinical Terms (formerly the Read Codes), SNOMED Clinical Terms.

The International Classification of Diseases (ICD) was first published in 1900. The basis was developed in 1839 by Farr.

5.1.2 Medical Patient Record

Comparable patient data are the key to improved effectiveness and efficiency in health care. The idealized reuse of clinical experience is highly dependent on consistent and comparable descriptions, the very purpose of clinical nomenclatures. The first effort for the above issue

was to adopt a standard derivative of the ICD for clinical use. The advent of ICD-9-CM involved the collaboration and cooperation of the major medical societies, associations, payers, government, and industry. Also, at the other end of the terminology spectrum regarding detailed clinical nomenclatures, a similar convergence and cooperation occurred. In a series of meetings sponsored by the Computer-based Patient Record Institute (CPRI), industry representatives of payers, providers, academia, and government began coalescing terminology principles on the path toward establishing comparable content. Three key observations have emerged from the meetings: a definition of clinical terminology, a recognition of a synergistic spectrum running from detailed nomenclatures to highly aggregating classifications, and the separation of thinking about terminologies into phases of use.

First, the definition for a clinical terminology is as follows:

> Standardized terms and their synonyms which record patient findings, circumstances, events, and interventions with sufficient detail to support, outcomes research, quality improvement; and can be efficiently mapped to broader classifications for administrative, regulatory, oversight, and fiscal requirements.

Second, the recognition that nomenclatures might complement and not compete with classifications resolves what has been a very long running controversy. It is self-evident that well-defined nomenclatures can be ”rolled-up” into aggregating classifications, although the rules and logic about how exactly to undertake this are not always obvious or explicit.

Third, the phases of terminology use are now widely regarded as entry terms, reference terminologies, and aggregate or administrative classifications. Entry terms are colloquial expressions or terms (in the strict sense of the word) that are familiar to users and convey sufficient specificity to say what is meant. These are translated into an underling reference terminology, which is capable of semantic closure and unambiguous representation. Finally, the formal reference terms can be aggregated using explicit inclusion, exclusion, and cross-referencing rules (which are not always readily available in machine readable form) into highlevel classifications like ICD-9-CM.

Without exception, the most extraordinary convergence is the ongoing effort to derive a new work from the rich combined content and structure of SNOMED RT and the U.K. NHS Clinical Terms, which will be called SNOMED Clinical Terms.

Convergence and openness have not been limited to terminology content. The venerable GALEN effort, centered in Manchester, England, is now open software, sharing the core Program code (http://www.opengalen.org/)

Finally, while redundant standard terminologies proliferated, so did standards organizations concerned about terminology issues. During the first 18 months, the newly formed ISO TC 215 on Health Informatics has established a Working Group (no.3) on health terminologies. Progress within this working group has fostered the emergence of standards about standards, such as meta-vocabulary, a foundation model for health terminology, good terminology development indicators, semantic links, and models for nursing adaptation of meta-standards.

For the future the greatest challenge remains the problem of semantic normalization or closure among the myriad ways in which a concept can be composed across and within systems. The problem becomes tightly coupled with that of reconciling the role of semantic representation in the information model of health systems with the expression of modified meaning in a terminology. An often-used example contrasts placing a diagnosis in a ”family history” field of a medical record with modifying that diagnosis with a ”family history” qualifier from the terminology . These are, of course conceptually identical, if expressively variant.

5.1.3 BAIK-information model

The role of classifications and thesauri in the general information flow of the medical domain were summarized by Giere in the ‘70s (Figure 1):

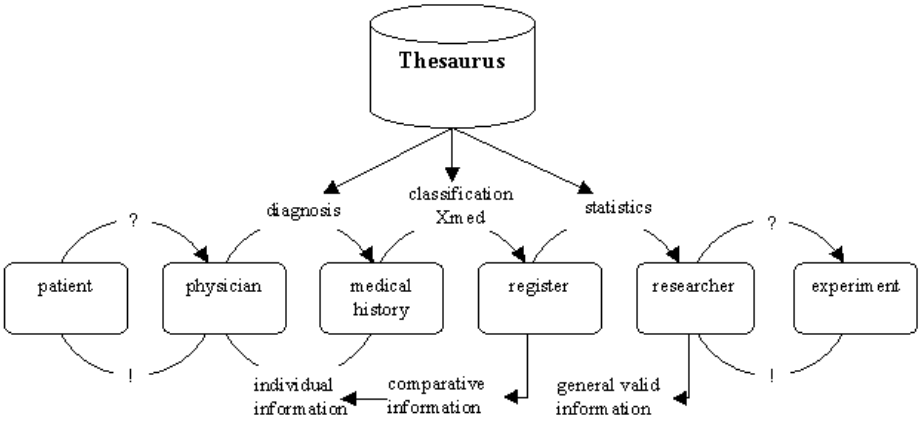

Figure 1: Information flow in medical practice and research

5.1.3.1 Theoretical basis

1. There is no classification in principle, every classification has a scope, just as information depends on the recipient and his situation.

2. There will be no new knowledge without standardization (i.e. controlled clinical trail, evidence based medicine)

but

3. Only with standardization there will be no new knowledge.

This means that substantially new concepts require unconventional thinking, impulses out of the practice, creativity, to step outside the predefined purposes. Like Albert Einstein an Isaac Newton, or the new concept in the medical domain, that the gastric ulcer is an infectious disease caused by helicobacter pylori. These would not be possible with a pre-established terminology. There must be space for language creativity. This is why Giere requires the

computers to follow the physician habits or physician creativity in language use and translate internally into standardized nomenclature rather than ask the physicians only to choose from pre-established nomenclature.

5.1.3.2 The concept of Patient Record

1. Primary patient record – open – ”translation”

   The primary patient record is the translation of the patient history. It is a collection of patient data acquired and filtered by the doctor. The doctor uses a variety of individual terms out of his everyday usage. He/she needs an open terminology for the primary patient record.

2. Secondary patient record – ”classification”

   This is the transformation of the primary patient record in a meta-patient-record by classification. The result should be ”language-free” or ”interlingual” preferably. Likes GALEN, SNOMED, UMLS (e.g. Moore with Med-parser.)

3. Tertiary patient record – (special) register

   The result of this first transformation, for example ICD, German OPS-§-301, is used for further analyses like medical statistics, epidemiology, reimbursement or, for administrative purpose.

5.1.3.3 Information cycles

1. Therapy-oriented information-cycle

   The patient comes to the doctor with a problem (?).The physician investigates the findings and enters them in the patient-record. This is the individual patient-oriented documentation. The doctor (or some other doctor) gets the individual information for the medical treatment, now or later (!).

2. Comparison-oriented information-cycle

   The patient findings, collected in the patient-orientated documentation, may meet classification for preset definitions and can be integrated in a standardized symptom-orientated documentation. This allows comparative information with other standardized cases, i.e. additional information for the treating doctor.

3. Knowledge-oriented information-cycle

   Statistical information is extractable out of the standard-documentation. The researcher will use this in order to formulate hypothesis (?), and model experiments so as to verify or reject the hypothesis (!). The extracted knowledge is the base for publications, textbooks, i.e. general valid information.

5.1.3.4 Terminology

A terminology represents the totality of terms used in one field of knowledge.

5.1.3.5 Thesauri (ontology)

A German standard of industry (DIN 1463) defines a thesaurus in an informational manner:

```
A thesaurus... is a list of  orderly terms and their
      names, which are used for indexing, storing, and
      retrieval in a special domain, with the following
      characteristics:

Terms and names are clearly related (terminological
      control) to each other by
      1. complete registration of synonyms
      2. defining one preferred term for every term
      3. representing the relationships between terms
```

A thesaurus is a list of interrelated terms used for a certain application area or domain. A thesaurus is always intended to be complete for its domain. For practical usage, thesauri that also contain a list of synonyms for each preferred term have also been developed (German ICD-10-thesaurus of diagnoses.) In this way, a thesaurus stimulates the usage of standardized terminology. Medical thesauri can be classified in:

Uni-dimensional thesauri: Only one dimension, like ICD

Multi-dimensional thesauri: Multiaxial thesauri like PCS, AGK-Thesaurus, ICD-10 thesaurus of diagnoses.

5.1.3.6 Controlled vocabularies

A controlled vocabulary is a restricted set of preferred terms used within an organization for a given purpose is called a controlled vocabulary.

5.1.3.7 Nomenclature

Nomenclatures are systematic organizations of names for describing objects in medicine. They represent a subset of terms out of a terminology (preferred terms.)

5.1.4 Documentation

5.1.4.1 Differences between the *acquisition* and *ordering of data* as medical documentation principles

The **acquisition** of data involves experience in medical practice and forms the implicit observational model which leads to the primary documentation of patient history.

The patient data can collected by the doctor, nurse or automatically from a laboratory information system, etc.

Acquisition of data is:

- unconscious, creative, open
- consistent, reproductible
- based on experience
- dialectic: Experience grows with collection

Data ordering with respect to a certain classification forms the explicit observational model and leads to the secondary documentation of patient history

Classification means sorting of data in classes according to specific criteria. Every classification is problem-orientated. There is no classification without a purpose.

Sorting requirements:

- Definition of purpose, one problem, deduced
- Non-overlapping class-system
- Defined criteria (standardization)
- Case-equity as a compromise of number of classes and their occupancy
- Refinement if new viewpoints appear within classification boundaries
- Reclassification if new viewpoints appear outside system boundaries

5.2 Medical Terminology: Coding and classification systems in health care

5.2.1 What are Clinical Terminologies?

Clinical Terminologies are defined lists of clinical terms or phrases, often with codes attached, whose purpose is to support the development of a clinical record that can be easily manipulated by computer systems. Examples are Read Codes and SNOMED Clinical Terms.

Clinical Terminologies should not be confused with *Classifications*. These are international standard coding schemes whose sole purpose is to collect one type of data (such as diseases) for statistical evaluation. They are not designed to create a complete, detailed clinical record. Examples of Classifications are ICD-10 and OPCS-4.2.

Simplifying, clinical terminologies are merely lists of words and phrases (”terms”) that are used in clinical practice: diseases, operations, treatments, drugs, administrative items, and so on. The following example is extracted from the ”Read Codes”:

Table 5: "Read Code" examples

Aachen aphasia test	Aarskog syndrome
Abacavir	Abacavir 20mg/mL oral solution
Abacavir 300mg tablet	Abachi wood RAST test
Abacteraemic sepsis	Abaete
Abalone canned in brine	Abandoned baby care
Abandoned child	Abandonment of elderly person
Abattoir fever	Abazinian language
Abbe anastomosis of jejunum to jejunum	Abbe flap
Abbe reconstruction of lip using distant flap of lip	Abbott Laboratories Ltd
Abbreviated eating attitudes test	Abciximab
Abdomen	Abdomen tympanitic

As the reader can see, the terms include all sorts of terms that may be required in order to describe the condition, circumstances or treatment of a patient.

5.2.2 What is the Purpose of Clinical Terminologies?

There are different types of Clinical Terminology, and they have different purposes. However, the three main scopes of a terminology are:

Table 6: Uses of clinical terminlogies

PURPOSE	EXAMPLES
Creating a computerized Clinical Record (Electronic Patient Record)	Read Codes, SNOMED
Summarizing the incidence of diseases and operations, on a national or worldwide level	ICD-9, ICD-10, OPCS-4
Managing the process of billing people for treatments they have received	CPT4, ICD-9CM

For the scope of this guide, the first group, which is the most complex, is considered to be the most interesting, and most useful goal of the Clinical Terminologies.

The main features of these Clinical Terminologies are:

- To enhance the development of computer systems which use clinical data
- To create a standard ”language of health” for use in healthcare computer systems
- To enable decision support to be performed by computer systems, such as checking whether a particular drug is suitable for a patient, given his medical history.
- To allow research and clinical management based on collected patient data

5.2.3 Synonyms in clinical terminologies

Clinical Terminologies usually have a way of representing synonyms words or phrases. For instance:

Table 7: Synonyms in clinical terminologies

TERM	SYNONYMS
Myocardial infarction	Heart attack Coronary thrombosis M.I. C.T.

To represent this, clinical terminologies usually have two types of code:

1. A *Concept* code, which is the same whatever synonym is used
2. A *Term* code, which is different for the different synonyms

For instance, the Read Concept and Term codes for the above example are as follows:

Table 8: Concept codes and term codes

TERM	CONCEPT CODE	TERM CODE
Myocardial infarction	X200E	Y202H
Heart attack	X200E	Ya0rZ
Coronary thrombosis	X200E	Ya0vv
M.I.	X200E	Ya0vw
C.T.	X200E	YaR89

One of the synonyms is usually identified as the *preferred*, or default, term, and in this example the preferred term is Myocardial infarction .

5.2.4 Concept Codes enable decision support and research

All the Concepts represented by the Terms in a Clinical Terminology are arranged in a hierarchy. The hierarchy describes the relationships between concepts. For example:

```
Myocardial infarction
is a type of
Ischaemic heart disease
which is a type of
Disorder of heart
which is a type of
Cardiovascular disorder
which is a type of
Disorder
which is a type of
Clinical finding
```

Now suppose that a doctor wishes to prescribe a drug that manifests side effects in patient with a certain heart disorder. Because the clinical terminology comprises every condition which is a type of heart disorder, the patient's record can automatically be checked to see whether the patient has any of these conditions.

This could not have been achieved with a free text patient record. For instance, Angina is a type of heart disorder, but this could not have been detected in a text-based patient record, where the best that can be achieved is to search for the word heart in the text. Nor would it have been possible to search for words heart OR angina OR coronary OR myocardial OR ... etc, as there are over 1000 types of heart disorder listed in the Read Codes for example.

One can now see how the hierarchy of concepts allows all sorts of research questions such as list all patients who have asthma, and list these according to the type of asthma that they have.

5.2.4.1 Coding for medical record abstraction

Computer-based patient data which are represented in a coded form have a variety of uses, including direct patient care, statistical reporting, automated decision support, and clinical research. There is no standard which supports all of these schemes. Several attempts were made, but none of the proposals have been widely accepted yet.

The coding of patient information has always been directed at simplifying the data, converting it to a general form which is easier to manipulate. Because the coding represents only a simplified synopsis of information extracted from the record, this kind of coding is referred to as abstraction. Record abstraction has been performed since the beginning of formal medical records, to allow assessment of incidence of a disease, mortality after surgical procedure or (in the era of prospective payment) costs evaluation for a hospital stay. The archetypical coding system for medical record abstraction is the International Classification of Diseases (ICD).

5.2.4.2 Coding for medical record systems

Abstracting systems are a fact of life for medical record keeping, both for health statistics reporting and for reimbursement. But they are not very useful for coding a research database, treatment decisions, case review, summary review, decision support, research, quality assurance and reporting of mortality and morbidity.

Electronic medical record (EMR) systems have the greatest vocabulary requirements. Standard vocabularies are mostly inappropriate for use in EMR, which motivates the development of controlled vocabularies. Some are based on a semantic network, like the Medical Entities Dictionary (MED) used in Columbia-Presbyterian clinical information system. This vocabulary integrates terms from national coding schemes with those from local ancillary systems to produce a unified coding scheme that retains the fine granularity from the original coding schemes while accommodating the coarser granularity of a variety of applications making use of the patient data. The semantic network model is useful both for supporting the addition of new terms from ancillary systems and for maintaining currency with changes in the national vocabularies.

With several decades of experience in computer-based vocabulary requirements, researchers are now beginning to collaborate to apply their individual experiences to the task of developing general purpose, comprehensive controlled vocabularies to support health care applications.

5.3 Documentation

5.3.1 Primary documentation

The primary documentation is the documentation in natural language by dictation or manual entry in a patient record. It is useful in extending terminology control of primary documentation and narrative text in order to find new terms and to expand the existing terminology. The primary documentation should be as structured as possible, but the doctor should not be restricted in his linguistic usage.

5.3.2 Secondary documentation

The secondary documentation is the translation of the primary patient record into a standardized meta patient record. This should be done automatically to the best possible extent.

5.3.2.1 SNOP: Standard Nomenclature of Pathology

One domain with a successful abstracting scheme is anatomic pathology. The college of American Pathologists developed the Standard Nomenclature of Pathology (SNOP) as a multiaxial system for describing pathologic findings through postcoordination of topographic (anatomic), morphologic, etiologic and functional terms.

SNOP was the basis for SNOMED, which includes even more axes. SNOMED aims at covering the complete medical terminology.

5.3.2.2 SNOMED: Systemized Nomenclature of Human and Veterinary Medicine

First published 1975 from the American Pathologists.

Multiaxial system (now eleven), each of these axes forms a complete hierarchical classification system. A diagnosis may consist of a topographic code, a morphology code, a living organism code, and a function code. When a well-defined diagnosis for a combination of these four codes exists, a dedicated diagnostic code is defined.

5.3.2.3 SNOMED-CT - SNOMED Clinical Terms

”SNOMED Clinical Terms” (or ”SNOMED-CT”) is the proposed name of a new clinical coding scheme, which is a merger of the Read Codes and SNOMED-RT. Its completion is currently scheduled for December 10, 2001.

5.3.2.4 SNOMED RT - Systematized Nomenclature of Medicine Reference Terminology

SNOMED RT is the recent version of SNOMED.

With over 340,000 explicit relationships, SNOMED RT will provide a common reference point for comparison and aggregation of data throughout the entire health care process.

More information on the SNOMED project can be found online at http://www.snomed.org/.

5.3.2.5 UMLS: The Unified Medical Language System

A long-term research and development project at the U.S. National Library of Medicine (NLM) since 1986 whose goal is to develop resources that will support intelligent information retrieval. The UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable biomedical information. The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people. The second is the distribution of useful information among many disparate databases and systems. A multidisciplinary developer team is working on the project. UMLS is a multilingual system. The Specialist Lexicon is being translated into German (DSL – Deutsches Specialist Lexikon, Zinfo-Frankfurt and Dep. Med-Inform. University of Freiburg.)

5.3.2.5.1 UMLS Metathesaurus

The Metathesaurus may be seen as a thesaurus that transcends individual thesauri (e.g., bibliographic databases, clinical record systems, expert systems), or controlled vocabularies and classifications. Entries in the Metathesaurus connect alternate names for the same concept, such as synonyms, lexical variants, and translations. Many attributes of individual concepts are also included. Some key attributes were created expressly for the Metathesaurus, others are taken from its source vocabularies. Metathesaurus entries may include multiple definitions from different sources, in which case each definition is labeled with its source. Special lexical entities such as acronyms, abbreviations, trade names, and drug identification numbers are explicitly labeled. All Strings are represented in the word index that accompanies the Metathesaurus. The index can be used to identify all concepts, terms and strings containing a particular word.

Table 9: Pneumonia concepts in the UMLS Metathesaurus

Bacterial Pneumonia
Pneumonia, Lobar
Pneumonia, Staphylococcal
Pneumonia, Streptococcal
Pneumonia due to Streptococcus
Pneumonia in anthrax
Bronchopneumonia
Pasteurellosis, Pneumonic
Salmonella Pneumonia
Pneumonia due to Klebsiella Pneumonia
Pneumonia due to other specified bacteria
Pneumonia in whooping cough
Pneumonia due to Pseudomonas
Pneumonia due to Hemophilus influenza (H. influenza)

5.3.2.5.2 UMLS Semantic Network

The UMLS Semantic Network provides a consistent view of the concepts represented in the UMLS Metathesaurus. Each concept in the Metathesaurus is assigned to one or more of the semantic types in the Network based on the meaning or meanings that the concept has in its source vocabularies. Assigning semantic types to Metathesaurus concepts involves algorithmic procedures as well as extensive review by subject matter experts. Wherever possible, default semantic types are assigned to concepts by a computer program. This is possible because most of the constituent vocabularies in the Metathesaurus are already structured, providing useful semantic information. These default assignments are subsequently reviewed by experts who determine if the correct assignment has been made and whether any types need to be added. The primary relation in the Semantic Network is the `isa` link. It links semantic types of greater and lower specificity, establishes the hierarchy of types within the Network, and is used for deciding on the most specific semantic types available for assignment to a Metathesaurus concept. The `isa` link allows nodes in a hierarchy to inherit information from higher level nodes. The non-hierarchical relationships in the Network fall into four categories: `physical`, `functional`, `temporal`, and `conceptual` relationships. The links indicate what relationships are possible (or permitted.) The Semantic Network importantly provides an overall semantic structure for Metathesaurus concepts. Since Metathesaurus concepts are derived from a number of thesauri which have their own structure, the Network exerts a unifying force. It groups together all concepts that share a particular semantic type and allows generalizations to be made about that set of objects.

5.3.2.5.3 UMLS Information Sources Map (ISM)

The Information Sources Map (ISM) is a knowledge source which describes computerized biomedical information sources. ISM records contain highly structured information, drawn in some cases from other UMLS Knowledge Sources, as well as information intended primarily for humans to read. The current version contains data on some 64 information sources. Four

elements in the ISM are used to index the conceptual scope of the information sources: relevant MeSH terms, MeSH subheadings which denote the contexts in which the main MeSH headings are applicable, semantic types from the UMLS Semantic Network, and semantic links, which link two semantic types with a relation from the Semantic Network.

More information on the UMLS project can be found online at http://www.nlm.nih.gov/research/umls/.

5.3.2.6 Specialist Lexicon

The SPECIALIST lexicon is an English language lexicon with many biomedical terms. It has been developed in the context of the SPECIALIST natural language processing project at NLM. The current version includes some 108,000 lexical records, with over 186,000 strings.

The lexical entry for each word or term records syntactic, morphological, and orthographic information. Lexical entries may be single or multi-word terms. Entries which share their base form and spelling variants, if any, are collected into a single lexical record. The base forms are the uninflected forms of the lexical item; that is, the singular form in the case of a noun, the infinitive form in the case of a verb, and the positive form in the case of an adjective or adverb.

Lexical information includes syntactic category, inflectional variation (e.g., singular and plural for nouns, the conjugations of verbs, the positive, comparative, and superlative for adjectives and adverbs), and allowable complementation patterns (i.e., the objects and other arguments that verbs, nouns, and adjectives can take.) The lexicon recognizes eleven syntactic categories or parts of speech: verbs, nouns, adjectives, adverbs, auxiliaries, modals, pronouns, prepositions, conjunctions, complementizers, and determiners.

The number and nature of the complements taken by verbs determine the basic sentence patterns of a language. The lexicon recognizes five broad complementation patterns: intransitive, transitive, ditransitive, linking and complex-transitive. Verb entries also encode each of the inflected forms (principal parts of the verb.) Verbs are inflectionally classified as regular, Greco-Latin regular or irregular. Noun entries describe the inflection of the nouns (pluralization) and spelling variations. Complementation patterns for nouns and nominalization information are also included where relevant. In addition to inflection and complement codes, adjectives in the lexicon have position codes to indicate the syntactic positions in which they may occur. An adjective may be a qualitative, classifying, or color adjective. Adverbs in the lexicon are coded to indicate their modification properties. The lexicon recognizes sentence, verb phrase and intensifier type adverbs, and classifies sentence and verb phrase adverbs into manner, temporal and locative types.

Lexical items are selected for coding from a variety of sources, including lexical items from MEDLINE® citation records, and a large set of lexical items from medical and general English dictionaries.

More information on the SPECIALIST project can be found online at http://www.nlm.nih.gov/nlmhome.html.

5.3.2.7 GALEN: Generalized Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine

5.3.2.7.1 The reason for the development of GALEN was mainly

the lack of a concept representation for medicine which is:

- Re-usable and application-independent
- Implementable
- Acceptable to:
    - Healthcare Professionals
    - Healthcare informaticians and computer scientists
    - System vendors

5.3.2.7.2 GALEN-CRM: (Common Reference Model)

is a formal representation of medical knowledge. It comprises:

- elementary clinical concepts such as fracture, bone, left, and humerus;
- relationships (e.g. as fractures can occur in bones), that control how concepts may be combined;
- complex concepts such as fracture of the left humerus composed from simpler ones.

5.3.2.8 GRAIL

Represants a formal language describing the rules for manipulating GALEN concepts and relationships, e.g., pathological fracture:

```
Fracture which
     <hasLocation-
      (AnatomicalNeck which
         isDivisionOf-(Femur which
               hasLaterality-right))
       hasCause-Osteoporosis>
```

5.3.2.9 GALEN – CRM using GRAIL formalism

is implemented in three modules which together form a Terminology Server.

5.3.2.9.1 Concept Module

The Concept Module (CM) and associated modeling tools allow terminology developers to create models containing concepts and relationships, and to derive new concepts that are valid compositions of existing ones. GRAIL allows the system to use the concepts and relationships to:

- determine whether or not a particular composition is sensible;
- generate all possible concepts based on that knowledge;
- automatically derive other relationships, such as classification hierarchies, based on the composition (definitions) of concepts. The number of unique medical expressions is 10^7.

In one domain (AIDS) there are : 150.000 candidate term phrases of 1 to 5 words each. GALEN comprises 100-200 medical subdomains, with an estimated 2-word expressions of $4*10^6$. This fact assumes 20.000 meaningful single words with a 10% combination rate.

5.3.2.9.2 Multilingual Module (MM)

GALEN separates the model of the concepts (*ideas*) from the natural language phrases used to refer to them (*terms*.) The reference model is intended to be language independent, so that information entered in one language can be displayed in another. The natural language phrases for concepts are generated by the **Multilingual Module (MM)** within the Terminology Server using the structure of the concept, and appropriate lexicons and grammar rules associated with the reference model. As a minimum, these lexicons must contain words for the elementary concepts. This makes the task of translating a terminology much smaller than that of translating all the possible terms. Phrases for complex compositions can be generated from these individual components by the Multilingual Module.

5.3.2.9.3 Conversion Module

The Code Conversion Module (CCM) Existing coding schemes are very important to GALEN. These schemes are widely used (and frequently mandatory) in current information systems and represent a large investment in expertise. Many schemes are detailed and aim for extensive clinical coverage. However, they typically lack the structure and formal basis that is necessary to meet the needs of advanced systems. GALEN relates to existing schemes by:

- drawing on existing schemes to help construct the reference model;
- mapping concepts in those schemes to structured concepts in the reference model;
- acting as an *interlingua* between schemes, thus supporting sophisticated code conversion;

- enhancing existing schemes by using the structure of the reference model to derive new relationships and verify or correct existing ones (e.g. classification hierarchies)

This functionality is the prime responsibility of the **Code Conversion Module (CCM)** within the Terminology Server.

5.3.2.10 The GALEN Terminology Server

The main modules (concept, multilingual, and code conversion) are integrated into a single multi-user, networked software system, the **GALEN Terminology Server (TeS.)** The TeS combines the functionality of the three modules to provided sophisticated but uniform terminology services to client applications. It embodies GALEN's view of terminologies as dynamic functional systems, rather than the traditional view as static data files. Client applications can pose high-level requests to the TeS, such as `what are the kinds of this` or, more interestingly, `what can I say about this`. The GALEN TeS represents a pervasive enabling technology for the electronic patient record by:

- supporting detailed clinical descriptions based on a semantically sound model of clinical terminology
- allowing arbitrarily complex clinical concepts to be stored in a fixed size representation for use in, for example, an existing patient record system;
- providing access to a powerful technology for structured data entry;
- offering sophisticated linguistic support allowing the rapid development of multilingual systems;
- facilitating the interchange of clinical data between systems which use different coding schemes and levels of clinical detail by offering a consistent view of coded data;
- preserving and adding value to what is already in existence by supplementing and extending existing coding and classification schemes

Table 10: Main differences between thesauri and formal classifications

THESAURI	FORMAL CLASSIFICATIONS
children narrower than parents mixture of kinds-of / parts	Children strictly kind of parents clean separation
human readable	machine readable
information is spread over structure and text (rubrics)	all information is in the structure
mono-hierarchical	dynamic reclassification
fixed	generative

Table 11: Medical language versus medical concepts

THE GALEN VIEW:	THE LE VIEW:
linguistic knowledge	phonologic knowledge
conceptual knowledge	morphologic knowledge
pragmatic knowledge	syntactic knowledge
criteria knowledge	semantic knowledge
terminological knowledge	pragmatic knowledge
	world knowledge

More information on *Open*GALEN can be found online at http://www.opengalen.com/.

5.3.2.11 Terminology server project at Zinfo

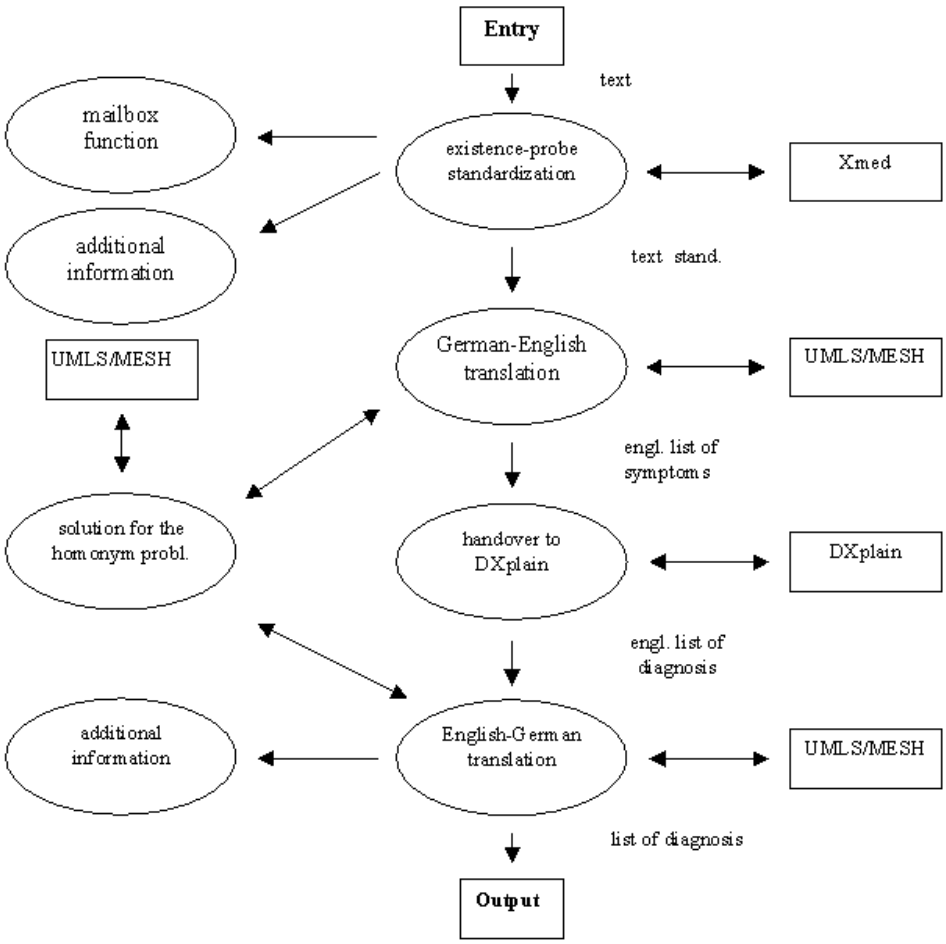

Figure 2: Terminology server project

The main goal of this project is to use an English expert-system with a crosslingual (in this case German) input. The picture above shows the workflow of the system.

5.3.2.12 AGK-Thesaurus

Is a thesaurus developed in a GMDS work group for plain-text documentation (GMDS stands for German Association of Medical Informatics, Biometrics, and Epidemiology - Arbeitsgruppe Klartextdokumentation der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V.)

It contains about 100.000 terms in the most recent version. The thesaurus organizes medical terms by semantic relationships. Synonyms are mapped to preferred terms, which are linked with other preferred terms.

AGK-Thesaurus has been in clinical routine for more than 25 years, particularly in various BAIK-systems.

An illustration of the thesaurus is shown in the following diagram:

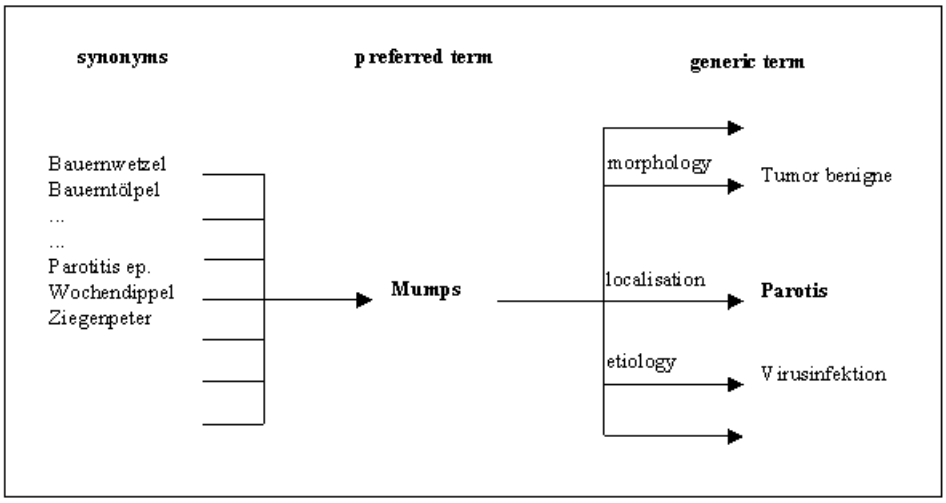

Figure 3: The preferred term `Mumps` and its possible links with the thesaurus

5.3.2.13 Read Codes

The Read Codes are a comprehensive list of terms intended for use by all healthcare professionals to describe the care and treatment of their patients. They enable the capture and retrieval of patient-centered information in natural clinical language within computer systems.

The Read Codes are used by a significant proportion of family practitioners in the UK to record details about patient care and for the business needs of the practice. Increasingly, the UK's acute and community healthcare sectors are using Read for recording patient-centered information and generating statutory returns to the UK Department of Health by use of the cross-references to international classifications such as the International Classification of Diseases (ICD-9 and ICD-10.)

The Read Codes are Crown Copyright and belong to the UK Department of Health. No charge is made for the intellectual property or significant development costs incurred.

The Read Codes are compiled and updated every six months for the full release and monthly for drugs. This work is coordinated by the NHS Information Authority working closely with all the clinical professions including doctors, nurses, professions allied to medicine, and pharmacists.

ReadEngine

The Lernout and Hauspie ***ReadEngine*** is a software toolkit which allows a developer to quickly and easily incorporate Read Codes, SNOMED Clinical Terms, or other clinical terminology into a new or existing healthcare system. Over 60 developers are currently using the ReadEngine to implement Read Codes.

5.3.2.14 ICD-10 thesaurus of diagnoses

is an easy to use tool for the medical practitioner to find the right code for a diagnose. In the published version it contains about 58.000 coded everyday use terms (including permutation of terms) in everyday use by the general practitioner. The software version contains 31.000 terms (not including permutations and synonyms) for easy retrieval.

5.3.2.15 Medical Classifications

Are used primarly in epidemiology for the statistical evaluation of mortality and morbidity), of populations. Futhermore they are used for reimbursement and administration purposes in the health care system.

5.3.2.15.1 International Classification of Diseases (ICD)

A classification of diseases can be defined as a system of categories to which morbid entities are assigned according to established criteria. The purpose of the ICD is to permit the systematic analysis, interpretation and comparison of mortality and morbidity data collected in different countries or areas and at different times. The ICD is used to group diagnoses of diseases and other health problems, and permits not only easy storage but also statistical presentation, retrieval and analysis of the data in a systematic way. In practice the ICD has become the international standard diagnostic classification for all general epidemiological and many health-management purposes.

Although the ICD is suitable for many different applications, it does not always allow the inclusion of sufficient detail for some specializations, and sometimes further information and

different attributes of the classified conditions may be needed. The main ICD (the three- and four-character classification) covered by the three volumes of the ICD-10 could not incorporate all this additional information while at the same time remaining accessible and relevant to its traditional users, so the idea arose of a ”family” of disease and health-related classifications including volumes published separately from the main ICD, taylored to specific requirements.

A statistical classification of diseases must be confined to a limited number of mutually exclusive categories and capable of encompassing the whole range of morbid conditions. The element of grouping distinguishes a statistical classification from a nomenclature, which must have a separate title for each known morbid condition. The concepts of classification and nomenclature are nevertheless closely related because a nomenclature is often arranged systematically.

A statistical classification allows for different levels of detail if it has a hierarchical structure with subdivisions. In addition, it should retain the ability both to identify specific disease entities and to allow statistical presentation of data for broader groups, to allow the obtaining of useful and understandable information. The ICD has developed as a practical, rather than a purely theoretical classification based on etiology, anatomical site, circumstances of onset, etc.

The ICD-10 uses an alphanumeric code with a letter in the first position and a number in the second through fourth positions. The fourth character follows a decimal point. This yields possible code numbers from A00.0 to Z99.9.

WHO (World Health Organisation) is responsible for its maintenance and

- to conduct the process of elaborating periodic versions
- to develop new methodologies for classifying and analysing data
- to facilitate training on ICD, its family and its framework in member countries
- to facilitate the improvement of the basic data

There are nine WHO Collaborating Centers for Classifications of Diseases, located in Australia, Brazil, China, France, Russia, Sweden, UK, USA and Venezuela, working as a network. The most recent version is ICD-10 and was published 1992. The U.S. National Center for Health Statistics published a set of clinical modifications to ICD-9, known as ICD-9-CM. Many countries publishes there own modifications, for instance Australia ICD-10-AM (Australian Modification.) which is hierarchically structured.

5.3.2.15.2 DSM: Manual of Mental Disorders

Another specialized coding scheme is the American Psychiatric Association’s Manual of Mental Disorders (DSM). Each edition of DSM has corresponding editions of ICD.

5.3.2.15.3 ICPM – International Classification of Procedures in Medicine

First published 1976 by WHO, the ICPM represented a source of inspiration for a number of other procedural classifications. The procedural part of ICD-9-CM was based on ICPM. OPS-

§-301 an extension of ICPM is mandatory in hospitals for reimbursement and administration purposes in Germany.

5.3.2.15.4 PCS – Procedure Coding System

Is a global procedure coding system, developed in the United States of America. The Health Care Financial Administration (HFCA) ordered the development of a new procedure classification system, because the procedure classification ICD-9-CM (in use since 1979) is expected to be insufficient in the future. 3M Health Information System (HIS) developed the new system ICD-10 Procedure Coding System.

PCS has a multi-axial, 7-character alphanumeric code structure. Each code character can have up to 34 different values (the ten digits 0-9 and the 24 letters A-H,J-N and P-Z). The letters O and I are not used in order to avoid confusion with the digits 0 and 1. Procedures are divided into sections that relate to the general type of procedure (e.g., medical and surgical, imaging, etc.) The first character of the procedure code always specifies the section. The second through seventh characters have a standard meaning within each section but may have different meanings across sections.

Prof. Dr. med. W. Giere edited the German translation of PCS on behalf of 3M Health Information Systems. The German Health Administration is evaluating the PCS as an option for a possible new procedure coding system.

5.3.3 Tertiary documentation

The tertiary documentation includes the abstraction of patient history metadata. This is mainly used in implementing the DRG system of reimbursement.

5.3.3.1 DRG: Diagnosis Related Groups

Are an American development for the purpose of abstracting medical records. Developed initially at Yale University (R-DRGs) for use in prospective payment in the Medicare program, themain scope of DRG coding is to provide a relatively small number of codes for classifying patient hospitalizations while at the same time providing some separation of cases based on severity of illness. One of the most recent DRG-systems is that developed in Australia 1998 (AR-DRGs, Australian Refined-DRGs.) This system will be adapted in the near future in Germany for reimbursement.

5.4 Medical Information Access

5.4.1 Institutions

5.4.1.1 DIMDI (German Institute of Medical Documentation and Information)

Through DIMDI, the German government supports many nomenclature activities. DIMDI is in charge of publishing German versions of official classifications ordered by the German Health Government. These include the International Classifications of Diseases (ICD-9, ICD-10), the procedure-coding system OPS-§ 301 SGB V and the Universal Medical Device Nomenclature System (UMDNS). In addition, the German translation of the Thesaurus Medical Subject Headings (MeSH) is provided by DIMDI and updated yearly.

These German Data (corpora) are free downloadable from the DIMDI-Server:

- ICD-9 (version 6)
- ICD-10 (version 1.3)
- OPS-§ 301 SGB V (version 1.1)
- UMDNS (version 1.1)

Also downloadable from DIMDI-Server are

- PCS: Procedure Coding System
- English-Version
- German-Translation (edited by Prof. Dr. med. W. Giere on behalf of 3M H.I.S.)

More information on DIMDI can be found online at http://www.dimdi.de/.

5.4.1.2 NLM (National Library of Medicine)

The National Library of Medicine is the world's largest medical library. The Library collects materials in all areas of biomedicine and health care, as well as works on biomedical aspects of technology, as well as other human, physical, life, and social sciences. The collection compromises more than 5.8 million items- books, journals, technical reports, manuscripts, microfilms, photographs and images. Housed within the Library is one of the world's finest medical history collections of old and rare medical works. The Library's collection may be consulted in the reading room or requested on interlibrary loan. NLM is a national resource for all U.S. health science libraries through a National Network of Libraries of Medicine®.

For more than 100 years, the Library has published the Index Medicus®, a monthly subject/author guide to articles in 3400 journals. This information, and much more, is today available in the database MEDLINE® via the World Wide Web.

More information on NLM can be found online at http://www.nlm.nih.gov/.

5.4.2 Systems

5.4.2.1 MeSH: Medical Subject Headings

This classification, too, is developed and maintained by the National Library of Medicine (NLM) in the United States. It is generally used to index the world medical literature. MeSH forms the basis of the Unified Medical Language System (UMLS) also developed by NLM. MeSH arranges terms in a structure that breaks from the strict hierarchy used by most other coding schemes. Terms are organized into hierarchies and may appear in multiple places in the hierarchy. Although it is not generally used as a direct coding scheme for patient information, it plays a central role in the Unified Medical Language System.

A German translation provided by DIMDI is updated every year. The following diagram illustrates a sample of the MeSH categories:

Table 12: MESH categories for `Pneumonia`

```
Respiratory Tract Diseases
     Lung Diseases
          Pneumonia
          Bronchopneumonia
Pneumonia, Aspiration
                    Pneumonia, Lipid
               Pneumonia, Lobar
               Pneumonia, Mycoplasma
               Pneumonia, Pneumocystis Carinii
               Pneumonia, Rickettsial
               Pneumonia, Staphylococcal
          Pneumonia, Viral
          Lung Diseases, Fungal
               Pneumonia, Pneumocystitis Carinii
     Respiratory Tract Infections
          Pneumonia
               Pneumonia, Lobar
               Pneumonia, Mycoplasma
               Pneumonia, Pneumocystitis Carinii
               Pneumonia, Rickettsial
               Pneumonia, Staphylococcal
          Pneumonia, Viral
          Lung Diseases, Fungal
               Pneumonia, Pneumocystitis Carinii
```

The list above shows a partial tree structure for the Medical Subject Headings (MeSH) showing pneumonia terms. Terms can appear in multiple locations, although they may not always have the same children, implying that they have somewhat different meanings in different contexts.

More information on MeSH can be found online at http://www.nlm.nih.gov/mesh/meshhome.html.

5.4.2.2 MEDLINE

MEDLINE® (Medical Literature, Analysis, and Retrieval System Online) is the U.S. National Library of Medicine's (NLM) premier bibliographic database that contains over 11 million references to journal articles in life sciences with a concentration on biomedicine. It has a time coverage from 1966 until present.

The sources are citations from 4,300 worldwide journals currently in 30 languages (40 languages for older journals cited back to 1966). About 52% of current cited articles are published in the U.S.; nearly 86% are published in English; about 76% have English abstracts written by authors of the articles. Citations for MEDLINE are created by the NLM, international partners, and cooperating professional organizations.

Weekly update: Approximately 8,000 completed references are added each Saturday, January through October (over 400,000 added per year.) Updates are irregular in November and December as NLM makes the transition to a new year of Medical Subject Headings (MeSH®) vocabulary used to index the articles.

MEDLINE has a broad coverage of the basic biomedical research and clinical sciences since 1966, including nursing, dentistry, veterinary medicine, pharmacy, allied health, and pre-clinical sciences. MEDLINE also covers life sciences that are vital to biomedical practitioners, researchers, and educators, including some aspects of biology, environmental science, marine biology, plant and animal science as well as biophysics and chemistry. Increased coverage of life sciences began in 2000.

MEDLINE is available on the Internet through the NLM home page at http://www.nlm.nih.gov/databases/freemedl.html and can be searched free of charge. No registration is required. MEDLINE services are also provided by organizations that lease the database from NLM. Access to various MEDLINE services is often available from medical libraries, many public libraries, and commercial sources as well as onsite at NLM in Bethesda, Maryland.

The MEDLINE link from the Welcome statement on NLM's home page leads to two web-based services, PubMed® and Internet Grateful Med, both offering MEDLINE search functionality. MEDLINE can be searched using NLM's controlled vocabulary MeSH, or by author name, title word, text word, journal name, phrase, or any combination of these. The result of a search is a list of citations (including authors, title, source, and often an abstract) to journal articles. Both of NLM's web-based search interfaces for MEDLINE also search MEDLINE in-process citations that are added daily, as well as some citations that come electronically directly from publishers. MEDLINE's in-process and the publisher supplied citations are not indexed with MeSH.

5.4.3 Tools

5.4.3.1 DXplain

DXplain is a diagnostic decision support program from Massachusetts General Hospital. It provides decision support for performing a differential diagnosis. DXplain is not intended to give the ”right answer,” it rather helps the physician during the process of diagnosing and gives warnings if something seems to be wrong.

Key components of Dxplain are:

Knowledge Base, which comprises a large Database storing findings (called terms, see below), diseases and relations between these two types.

It is designed to provide plausible explanations for a given set of signs and symptoms, having additional text for each disease, which describes the nosological entity and contains literature references.

Entities are arranged in hierarchies. Important for the user is the hierarchical interface as well as the rating algorithm.

Terms and Diseases:

Terms in DXplain represents medical findings.

Each term is assigned a value (range 1 to 5), called *term importance*, which is disease independent and describes the significance in defining a pathological condition. Low values indicate unspecific findings which describe with high probability healthy people. At the opposite end, high values are assigned to high reliability pathological findings which are rarely found in healthy subjects.

Like Terms, every disease holds a value (range 1 to 5), called *disease importance*, which indicates the seriousness of this condition. The disease importance is not used by the rating algorithm, but generares warnings in the presence of dangerous pathological situations.

The disease description is

composed of several terms and their significance index . The significance index specifies the intersection between terms and diseases and consists of two parts: *term frequency (disease to term relation)* and *evoking power (term to disease relation.)*

The rating algorithm

generates a list of possible diagnosis that match the given pathological or normal findings. It consists of two steps:

1. Selection

   From the domain of all existing diseases, the selection algorithm filters those that are

sufficient to explain the given findings. The rating algorithm generates several buckets. The first one represents a container for diseases, that could not be specified with high significance. These are diseases, for which at least one appropriate term exists. The last one contains diseases, that could be determined with high reliability, i.e. all terms match.

The containers between these two are arranged respectively.

Appropriate to the matching terms and their term importance , the diseases are stored to the suitable container.

The list of diagnosis which will be presented to the physician is generated out of these containers.

2. Rating

   The rating algorithm gives scores to the diseases delivered by the selection algorithm. The scores are calculated from the *term importance* and *evoking power*.

   The output of the rating algorithm are two sorted lists, one for common and one for rare diseases. Two leading positive signs (++) signalize diseases, that could directly derive from the given findings. A diseases marked with the symbol * needs immediate therapy, in this case diseases importance is used.

The user Interface

DXplain is designed for easy and interactive use. Important features are:

- Usage of key words and synonyms during input
- Automatic recognition and correction of misspelled words
- Menu-driven user interface
- On-line help
- Various functions for additional explanation, e.g. explanation of differences between two diseases.
- Options for influencing the rating algorithm, e.g. focus on term causes the algorithm to consider only diseases, which have this term.

Summary:

DXplain is a diagnostic decision support system, designed for supporting physicians during differential diagnosis. It uses a large knowledge base, consisting of findings (terms), diseases and relations between these two sets. It uses a probability-based rating algorithm and has

various explanation capabilities. DXplain should be used like a medical book; it does not replace the physician's expert knowledge.

More information on DXplain is available online at
http://www.cpmc.columbia.edu/homepages/ciminoj/present/acmi96/dxplain.htm.

5.4.3.2 Xmed

Is an application for plaintext processing of medical text and classification according ICD and OPS § 301 SGB V.

Xmed is a modular computer-system for classifying and standardizing medical text of diagnoses and therapies. Data imported from any source by SGML transformation thus allowing a semantic description. Xmed makes use of TRANSOFT (a thesaurus-based translation-system) in structuring the original data. The content derivation of the standardized data enables a coding with ICD and OPS § 301 SGB V (IKPM.) The results are exported in the desired format.

Client-server architecture are programmed in "Open M," which allows a portability to other operating systems.

Standardization of medical texts makes use of the efficient translation-system TRANSOFT (G.W. Moore, John Hopkins Medical Institute, Baltimore). The main features of the application developed at Zinfo are:

Text content derivation by use of a multiaxial medical thesaurus with more then 80.000 items (AGK-Thesaurus)

Rule-based, reproducible classification of diagnosis with ICD and OPS § 301 SGB V (IKPM)

Adaptive and extensible knowledge-base, even by the user

Stand-alone usability (for instance for coding)

Embeddable in other clinical information systems

5.4.3.3 TRANSOFT

Is a table driven German to English medical document translation system written in ANSI Mumps programming language, which allows an automated translation of German to English medical text.

A lexicon of words and idioms is one of two external tables of language-specific control information used by the TRANSOFT system. The lexicon consists of all acceptable source language words and idioms, their part of speech designators, and their primary and any alternative definitions.

A parsing table of word rearrangement instructions, or parsing formulas, is the second translation table used by TRANSOFT. Parsing formulas are applied recursively by TRANSOFT to transform a sentence in German word order (source) to its corresponding

English word order (target), after which English-to-German word and idiom substitution is performed.

5.4.3.4 Med-parser (Moore)

A prototype medical parser that tests sentences in routine medical text, especially in anatomic pathology reports. The diagnostic information in anatomic pathology reports exists predominantly in free-text. In order to recover this information for data-mining applications, the free-text must be computer-translated, or autocoded, into standardized medical coding languages, such as SNOMED, Read, or UMLS. As a first step in autocoding, each sentence in a pathology report must be a grammatically well-formed sentence, so that the autocoder can correctly identify critical elements, such as bodysite, diagnosis, negation, etc. and their relationships to on another. The MEDPARSE parser consists of a lexicon, parsing table, and a parsing script.

The allowable parts-of-speech for MEDPARSE are assigned according to the UMLS Specialist Lexicon.

MEDPARSE works on the principle of Reverse Backus Naur Form. The programming language is Perl.

5.4.3.5 MEDLEE: MEDical Language Extraction and Encoding System - Medical Language Processor

MedLEE is a Natural Language Processing application developed and in production at The New York Presbyterian Hospital. It was designed in the early 1990s to automatically encode text reports from the Department of Radiology. Most of the information in the medical domain is encoded as free text. Though useful for human readers it is difficult, if not impossible for databases to utilize the information effectively. MedLEE has four main components:

1. The Preprocessor

   The preprocessor uses ”report grammars” to divide the document into section.Words are matched to the semantic lexicon and coded with their semantic type. Irrelevant ”stop words” and known irrelevant phrases are removed from the text.

2. The parser

   MedLEE has a set of ”sentence grammars.” These are common patterns of semantic types which appear in radiology texts. The parser classifies groups of words based on these grammars. The grammars are predominantly semantic, though some syntactic information is used. If the sentence matches a grammar rule, a frame for the information is generated. If the sentence does not parse, MedLEE will try to reparse smaller segments of the sentence.

3. The Phrase Regularizer

   The semantic representation of the sentence is finalized. Uses a series of mapping rules to resolve split patterns. Split patterns occur when multi-word phrases are discontinuous in the text.

4. The Encoder

   Identified semantic units are matched to the Medical Entities Dictionary (MED), a controlled vocabulary in the CIS. The MED contains preferred terms for concepts and is used to reduce redundancy and ambiguity in the case of synonymy.

The MED limits the degrees of qualifiers. A final form of the CIS database is prepared.

5.4.3.6 MEDTAG: Tag-like Semantics for Medical Document Indexing

Is a project founded by the Swiss government (FNRS – Swiss National Foundation for Research) with the purpose to construct a semantic tagset for medical document indexing.

The UMLS hierarchical classes were used as a basis for the tagset.

Natural language processing seems to be the best way to handle a large amount of textual information, like in the medical domain.

The probabilistic approach was chosen for two reasons. First, for development time: HMM (Hidden Markov Model) taggers are data-driven and known to be easy to train. Second, for ignorance of semantic rules: unlike syntax, semantic rules and heuristics have not been deeply explored yet.

The corpora were operation reports from the abdominal surgery domain with large part of free text.

5.4.3.7 MedIAS Web Service

is a context-sensitive Web-Agent for internet and intranet use which retrieves updated, complementary, profile-oriented information for the EPR from a variety of filtered Web-sources. MedIAS analyses the thematic profile and medical context using Xmed, a thesaurus-based application, starting a web search routine. The overall findings of the search can then be filtered and presented to the physician in a profile-oriented, dynamically generated SGML document to enhance decision support. TCP/IP Sources used by the MedIAS prototype for data retrieval are: Dr-Antonius – a German medical web-crawler -, Medline and other medical databases over The German Institute for medical Information and Documentation (DIMDI), but the embedded use of local and CD databases is also possible. The future development of the project includes the retrieval of information from medical newsgroups, from guidelines of

the Scientific Forums of the Medical Specialist Associations (AWMF) and, last but not least from the expert system DXplain.

5.4.3.8 Dr. Antonius

is a German medical web-crawler with an integrated, intelligent medical dictionary. Search results are exclusively German web-pages even when search terms are language homonyms, thus overcoming the disadvantages of finding too many and too unspecific documents. The Web-robot with a specific medical list as a starting point, contacts medical websites as a background process, analyses their contents and stores it in a database that is consequently compared to a thesaurus. This allows a finer granularity of the search process. The recognition of German web-documents as such is done through prepositions whereas that of medical content through a carefully selected concept list including over 20,000 common medical terms. A web page is categorized as having a specific medical content only when a defined percentage of its terms are medical. The originality of Dr. Antonius lies in the thesaurus enhanced search option. The German ICD10 Diagnosis Thesaurus and the Xmed Thesaurus build a consistent backbone of the search-engine. Apart from a simple search, Dr. Antonius has also optional advanced search using the AND, OR, NOT or NEAR operators. The present database comprises over 65,000 German web pages with medical content.

5.4.4 Comparing Public Resources for a Medical Ontology

The perfect medical ontology is not available and probably will not be available in the immediate future, so this report explores the main terminological systems: SNOMED, GALEN, and UMLS. Other general semantic approaches have also been considered, but not presented because of relevance reasons. The items of ICD, and of classifications in general, do not necessarily correspond to the items found within texts, as they try to group them in classes. Futhermore, such entities are too complex (multi-word phrases) for the purpose of extracting a medical ontology.

SNOMED would be an extremely interesting source for indexing. Tags could be selected at a higher level within each of the 11 axis, although the links between the items are formalized to a very limited extent. But the content of SNOMED is limited to the medical domain and, as such, does not provide tags for general vocabulary. Nevertheless, SNOMED remains an interesting source to be considered, especially when comparing the content coverage of major clinical classifications with the content of patient records. Although SNOMED (nomenclature, but designated as classification in this comparison) obtained excellent references, some recent studies showed that UMLS had a better content coverage.

The GALEN project aims at developing a concept reference (CORE) model of medical concepts. It represents the concepts used in medical records or referred to by other coding systems or nomenclatures. Formally, the GALEN CORE model would provide the best basis for the building a multilingual medical ontology, as the concepts can in addition be annotated with words or terms in several languages.

These annotations allow concepts to be found via the lexicon entries. The hierarchy would allow tags to be defined at a more general level, and indexes to be attributed at the most detailed level. The hierarchies being multiple, it would be possible to find several aspects of a

same concept. At the same time, all the real ambiguities could be found, if a same annotation in a language is available for several distinct concepts.

However, a fundamental problem with using GALEN is its limited domain coverage: about 13000 concepts are contained in the model today, and few were relevant for the texts (abdominal surgery reports.) One major disadvantage of GALEN compared to UMLS or SNOMED is that the future and the maintaining of the model is unknown.

The semantic types of UMLS may be considered as a basic ontology for the domain, as these are quite general and allow the tagset to be limited. The current version of the semantic network contains around 130 classes, and around 50 dyadic relationships. Every entry of the metathesaurus is attributed to one or several classes. Although this network is sometimes regarded as being too general for medical purposes, it seemed to be at the right level for the purpose of elaborating an ontology.

A lexicon with semantic tag-like features and a probabilistic tagger to process the tag-like information were built. Whereas semantic tagging results open new perspectives in Medical Language Processing, mastering further semantic disambiguation may require more adapted tools, likely to cope with long and very long (out of the sentence) distance dependencies, similar to what is done in semantic clustering. Another problem arises from the maintaining of the probabilistic tagger. As biases may have important negative side effects, a simple rule-based assistant could improve performances significantly. Therefore some patterns extracted by the tagger once expressed in a symbolic formalism could serve as a basis for a future semantic rule-based tagger.

5.4.5 Data Mining

5.4.5.1 Overview

Generally our capabilities of generating and collecting data have been increasing rapidly in the last years. The computerization of many business transactions and the advances in data collection tools has provided us with large amounts of data. Especially in the medical domain it is necessary to generate and store huge amounts of various data.

This explosive growth in data and databases includes to an urgent need for new techniques and tools for retrieval. Tools for intelligent and automatic transformation of the processed data into useful information and knowledge are needed. The traditional manual data analysis has become insufficient, and methods for efficient computer-assisted analysis indispensable.

Therefore data mining, also referred to as knowledge discovery, has become a research area of increasing importance.

Data mining is defined as the process of nontrivial extraction of implicit, so far unknown and potentially useful information from data in databases. Many researchers have recognized mining information and knowledge from large databases as a key research topic. Moreover, several emerging applications in information providing services like the World Wide Web also call for several data mining techniques.

5.4.5.2 Scope in the medical domain

With the widespread use of medical information systems using databases, the medical data featured an explosive growth in size. Physicians and medical researchers are faced with the problem of making use of the stored data.

The general goal of applying data mining techniques in medicine is to improve directly or indirectly the quality of health care. Specific goals are the extraction of medical knowledge for diagnosis, screening, monitoring, research, therapy support and overall patient management. Some topics that are relevant in this special context of data mining are:

- Data mining techniques, particularly suited for medical applications
- Criteria for selecting specifical data mining techniques
- Quality assessment measures for data mining, e.g., validity, utility, comprehensibility, and novelty of discovered knowledge.
- Issues related to the representation of extracted knowledge.
- The integration of data mining tools into the existing medical information systems.
- Inclusion of medical experts/physicians in the preparation of data for data mining (e.g., data representation, modeling, cleaning, selection, and transformation), as well as in the interpretation of results.
- Distribution of results: How did the results of data mining affect medical practice or how did they assist in medical research.

5.5 Conclusion

Basically, the main problem of the medical domain is to get the right information for specific patient cases. Therefore several efforts were directed in the development of medical information systems, which are capable to answer domain specific queries and perform decisionally valid information retrieval. The main disadvantage of these methods is that they are only able to access information stored on the basis of rational considerations or proved assumptions. There is, however, information whose existence could not derived from these assumptions. Data mining offers help in finding this hidden information using an explorative approach: starting from the data itself, information hypotheses are generated and rated. With this strategy so far unknown occurencies can be provided, because irrational or redundant information is also a part of knowledge.

6 References

Abney, S. 1991. Parsing by Chunks. In *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.

ACL, 1997. Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization. Madrid, Spain.

Adriaens, G., and S.L. Small. 1988. Word expert revisited in a cognitive science perspective. In Small, S., G.W. Cottrell, and M.K. Tanenhaus (eds.): *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence.* Morgan Kaufmann, San Mateo, CA, pages 13-43.

Agirre, E., and G. Rigau. 1996. Word Sense Disambiguation using Conceptual Density. In *Proceedings of COLING96*, Copenhagen, Denmark.

Agirre, E., G. Rigau, L. Padró, and J. Atserias. 2000. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and the Humanities* 34(1/2):103-108.

Ahmad, K. 1994. Language Engineering and the Processing of Specialist Terminology. Presented at: *The Language Engineering Convention/Journees du Genie Linguistique*. Paris (France.) [http://www.computing.surrey.ac.uk/ai/pointer/paris.html; accessed 09/28/2000]

Al-Onaizan, Y., , J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N.A. Smith, and D. Yarowsky. 1999. Statistical machine translation, final report. JHU Workshop 1999. Technical report, CLSP, Johns Hopkins University.

Al-Onaizan, Y., U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, and K. Yamada. 2000. *Translating with Scarce Resources*. [http://www.isi.edu/~marcu/papers/aaai-00-tetun.ps; accessed 09/28/2000]

Alshwede, T. 1993. Word sense disambiguation by human informants. In *Proceedings of the Sixth Midwest Artificial Intelligence and Cognitive Society Conference*, Carbondale, IL, pages 73-78.

Amsler, R. 1980. The Structure of the Merriam-Webster Pocket Dictionary.PhD Dissertation, Univ. of Texas at Austin.

Arppe, A. 1995. Term Extraction from Unrestricted Text. In *Proceedings of the 10th Nordic Conference on Computational Linguistics*, Helsinki, Finland. [http://www.lingsoft.fi/doc/nptool/term-extraction.html; accessed 09/28/2000]

Atkins, S. 1993. Tools for computer-aided lexicography: The Hector project. In *Papers in Computational Lexicography (Proceedings of COMPLEX '93)*, Budapest.

Baker, C.F., C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Baldwin, B., and T.S. Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, Granada, Spain.

Baldwin, B., T.S. Morton, and A. Bagga. 1999. Overview of the University of Pennsylvania's Tipster Report. *In TIPSTER Text Phase III Proceedings October 96-October 98*, pages 151-162. Omnipress, Inc.

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

Ballesteros, L., and B. Croft. 1998 Resolving ambiguity for cross-language retrieval. In *21st Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 64-71.

Barnett, G.O., J.J. Cimino, J.A. Hupp, and E.P. Hoffner. 1987. DXplain. An evolving diagnostic decision-support system. *Jama* 258(1):67-74.

Barnett, G.O., K.T. Famiglietti, R.J. Kim, E.P. Hoffner, and M.J. Feldman. 1998. DXplain on the Internet. In *Proc AMIA Symp*, pages 607-611.

Barzilay, R., and M. Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, ACL, Madrid, Spain, pages 10-17.

van Bemmel, J.H. and M.A. Musen (eds.) 1997. *Handbook of Medical Informatics.* Bohn Stafleu van Loghum & Springer Verlag, Houten, NL. [http://www.mihandbook.stanford.edu/handbook/home.htm; accessed 09/28/2000]

Berry, M.W. 1992. Large-scale sparse singular value computations. *The International Journal of Supercomputer Applications*, 6(1):13-49.

Boguraev, B., and T. Briscoe (eds.) 1989. *Computational lexicography for natural language processing.* London: Longman.

Bourigault, D. *et al*. 1996. LEXTER: A Natural Language Tool for Terminology Extraction. In *Proceedings of the 7th International Congress of EURALEX*.

Boutilier, C., N. Friedman, M. Goldszmidt, and D. Koller. 1996. Context-specific independence in Bayesian networks. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI-96)*, Portland, Or.

Brants, T., and W. Skut. 1998. Automation of Treebank Annotation. In: *Proceedings of the Conference on New Methods in Language Processing (NeMLaP-3)*, Australia.

Braschler, M., and P. Schäuble. 1998. Multilingual Information Retrieval Based on Document Alignment Techniques. In Nikolaou, C., and C. Stephanidis (Eds.): *Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL '98*, Heraklion, Crete, Greece, September 21-23, 1998, Proceedings. Vol. 1513 of Lecture Notes in Computer Science, Springer Verlag, 1998, pages 183-197.

Braschler, M., Krause, J., Peters, C., and Schäuble, P. 1999. Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7).* NIST, Gaithersburg, MD.

Braschler, M., Peters, C., and Schäuble, P. 2000. Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8).* NIST, Gaithersburg, MD.

Braschler, M., Harman, D., Hess, M., Kluck, M., Peters, C., and Schäuble, P. 2000. The Evaluation of Systems for Cross-Language Information Retrieval. In *Proceedings of the 2nd International Conference on Language Resources & Evaluation (LREC 2000).*

Brown, P.F., J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, R.L. Mercer, and P.S. Roossin. 1988. A Statistical Approach to Language Translation. in *Proceedings of the 12th International Conference on Computational Linguistics,* Budapest, Hungary.

Brown, P.F., J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79-85.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical Machine Translation: Parameter estimation. *Computational Linguistics*, 32:263-311. [http://www.clsp.jhu.edu/ws99/projects/mt/ibm-paper.ps; accessed 09/28/2000]

Brown, R.D. 1996. The Pangloss-Lite Machine Translation System. in Expanding MT Horizons: *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*. Montreal, Canada. pp 268-272. [http://www.cs.cmu.edu/~ralf/papers.html; accessed 09/28/2000]

Brown, R.D. 1997. Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. In *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*. Santa Fe, July 23-25, pages 111-118. [http://www.cs.cmu.edu/~ralf/papers.html; accessed 09/28/2000]

Brown, R.D. 1998. Automatically-Extracted Thesauri for Cross-Language IR: When Better is Worse. In *Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98)*.

Brown, R.D. 1999. Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the Eighth International Conference on Theoretical and*

Methodological Issues in Machine Translation. Chester, England, August, pages 22-32. [http://www.cs.cmu.edu/~ralf/papers.html; accessed 09/28/2000]

Brown, R.D. 2000. Automated Generalization of Translation Examples. To appear in *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*. Saarbrucken, Germany. August. [http://www.cs.cmu.edu/~ralf/papers.html; accessed 09/28/2000]

Bruce, R., and J. Wiebe. 1994a. A new approach to word sense disambiguation. . In *Proceedings of the ARPA Workshop on Human Language Techonology*, Morgan Kaufman: San Francisco, CA.

Bruce, R., and J. Wiebe. 1994b. Word-sense disambiguation using decomposable models. In *Proceedings of ACL 32*, pages 139-145.

Bruce, R., and J. Wiebe. 1999. Decomposable models in natural language processing. *Computational Linguistics* 25(2):195-207.

Bruce, R., J. Wiebe, and T. Pedersen. 1996. The measure of a model. In *Proceedings of EMNLP 96*, pages 101-112.

Buchholz, S., J. Veenstra and W. Daelemans. 1999. Cascaded Grammatical Relation Assignment. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Baltimore, MD, pages 239-246.

Buckley, C., and Voorhees, E. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Buitelaar, P. 1998. *CoreLex: Systematic Polysemy and Underspecification.* PhD Dissertation , Brandeis University.

Buitelaar, P. 2000. Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification. In *Proceedings of ANLP2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems* , Seattle, WA.

Burgess, C., and K. Lund. 1997. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12(2/3):177-210.

Califf, M.E., and R. Mooney. 1999. Relational Learning of Pattern-Match Rules for Information Extraction. In *Proceedings of the 16th National Conference on AI (AAAI-99)*.

Carbonell, J.G., and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR-98*, Melbourne, Australia.

Carbonell, J.G., Y. Yang, R.E. Frederking, R. Brown, Y. Geng, and D. Lee. 1997. Translingual information retrieval: A comparative evaluation. In *Proceedings of IJCAI-97*, Nagoya, Japan.

Carl, M. 1999. Inducing Translation Templates for Example-Based Machine Translation. In *MT-Summit VII*.

Carl, M. 2000. Towards a Model of Competence for Corpus-Based Machine Translation. In *IAI Working Paper No. 36.* [http://rockey.iis.sinica.edu.tw/oliver/iaiwp/p8/; accessed 09/28/2000]

Carl, M. and S. Hansen. 1999. Linking Translation Memories with Example-Based Machine Translation. In *Proceedings of MT-Summit VII. Singapore*. [http://rockey.iis.sinica.edu.tw/oliver/iaiwp/p7/; accessed 09/28/2000]

Carl, M., C. Pease, and O. Streiter. 1999a. Examples of hybrid Machine Translation. In *ISMT and CLIP*, Beijing.

Carl, M., L.L. Iomdin, C. Pease, and O. Streiter. 1999b. Towards a Dynamic Linkage of Example-Based and Rule-Based Machine Translation. [http://rockey.iis.sinica.edu.tw/oliver/publ/dynlink2/dynlink2.html; accessed 09/28/2000]

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2):249-254.

Catizone, R., G. Russell and S. Warwick. 1993. Deriving Translation Data from Bilingual Texts. In *Proceedings of the First International Lexical Acquisition Workshop*. Detroit, MI.

Charniak, E. 1981. The Case-Slot Identity Theory. *Cognitive Science* 5.

Chaudiron, S., and Schmitt L. 2000. Amaryllis: An evaluation-based program for text retrieval in French. *In Proceedings of the 2nd International Conference on Language Resources & Evaluation (LREC 2000).*

Chen, S. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, Columbus, OH, pages 9-16.

Chen, H., and S.T. Dumais. 2000. Bringing order to the web: Automatically categorizing search results. In *Proceedings of CHI'00, Human Factors in Computing Systems*.

Chen, K.-H., and H.-H. Chen. 1995. Machine Translation: An Integrated Approach. In *Proceedings of TMI95*. Leuven, Belgium, pages 287-294.

Chodorow, M., C. Leacock, and G.A. Miller. 2000. A topical/local classifier for word sense identification. *Computers and the Humanities* 34(1/2):115-120.

Choi, S.-K., H.-M. Jung, C.-M. Sim, T. Kim, D.-I. Park, J.-S. Park, and K.-S. Choi. 1998. Hybrid Approaches to Improvement of Translation Quality in Web-based English-Korean Machine Translation. In *Proceedings of COLING-ACL*. Montreal, Canada.

Church, K.W., and P. Hanks. 1989. Word assiciation norms, mutual information, and lexicography. In *Proceedings of ACL 27*, Vancouver, Canada, pages 76-83.

Cimino, J.J. 1996. Review paper: Coding systems in health care. *Methods Inf Med* 35(4/5), pages 273-284.

Cleverdon, C. W., and Mills, J. (1963): The testing of index language devices. In *Aslib Proceedings* 15(4), pages 106-130.

Collier, A. *et al*. 1998. Refining the Automatic Identification of Conceptual Relations in Large-scale Corpora. In *Proceedings of the Sixth Workshop on Very Large Corpora*, ACL-COLING '98, pages 76-84.

Collins, B. 1999. Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach. Ph.D. thesis, Trinity College, Dublin.

Copestake, A. 1992. *The Representation of Lexical Semantic Information*. Doctoral dissertation, The University of Sussex. [= Cognitive Science Research Paper CSRP 280, 1992.]

Coret, A., Kremer, P., Landi, B., Schibler, D., Schmitt, L., and Viscogliosi, N. 1997. Accès à l'information textuelle en français: Le cycle exploratoire Amaryllis. In *Actes 1ères Jst Francil 1997* (in French.)

Cranias, L., H. Papageorgiou, and S. Peperidis. 1994. A matching technique in example-based machine translation. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, Kyoto, Japan, pages 100-104.

Craven, M., and J. Kumlien. 1999. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB-99).*

Cruse, D.A. 1986. *Lexical Semantics*. Cambridge Textbooks in Linguistics, Cambridge University Press

Cutting, D.R., J.O. Pedersen, D. Karger, and J.W. Tukey. 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR '92*, Copenhagen, Denmark, pages 318-329.

Daelemans, W., J. Zavrel, K. van der Sloot and A. van den Bosch. 1998. MBT: Tilburg Memory Based Learner, Version 1.0, Reference Manual. Technical Report ILK-9803, ILK, Tilburg University.

Dagan, I., L. Lee, and F. Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of ACL 35*, pages 56-63

Daille, B. 1994. Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques. PhD dissertation, Université Paris VII.

David, S., and P. Plante. 1990. De la nécessité d'une approche morphosyntaxique en analyze de textes. *Intelligence Artificielle et Sciences Cognitives au Québec* 3(3), 140-155.

Davis, M. 1996. New experiments in cross-language text retrieval at NMSUs Computing Research Lab. In *Proceedings of the Fifth Text Retrieval Conference (TREC-5).*

Davis, M., and W. Ogden. 1997. Quilt: Implementing a large-scale cross-language text retrieval system. In *20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97),* pages 92-98.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6):391-407.

Deutsches Institut für Medizinische Dokumentation und Information (DIMDI) (ed.) 2000. ICD-10-Diagnosenthesaurus. Sammlung von Krankheitsbegriffen im Deutschen Sprachraum, Verschluesselt nach der Internationalen Statistischen Klassifikation der Krankheiten und Verwandter Gesundheitsprobleme (ICD-10-SGBV, Version 1.3 [JULI 1999]) Version 3.0, Stand Januar 2000. Dreiländerausgabe Deutschland-Schweiz-Österreich. Hans Huber, Bern.

Dolan, 1994. Word Sense Ambiguation: Clustering related senses. In *Proceedings of COLING94*, pages 712-716.

Draper, S. 1998. Mizzaro's framework for relevance. [http://www.psy.gla.ac.uk/~steve/stefano.html; accessed 09/28/2000]

Duda, R.O., and P. E. Hart. 1973 *Pattern Classification and Scene Analysis.* New York: Wiley.

Dumais, S.T., and H. Chen. 2000. Hierarchical classification of web content. In Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00).

Dumais, S.T., T.K. Landauer, and M.L. Littman. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR'96 Workshop On Cross-Linguistic Information Retrieval*.

Eichmann, D., M.E. Ruiz, and P. Srinivasan. 1998. Cross-language information retrieval with the umls metathesaurus. *In 21st Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98),* pages 72-80.

van der Eijck, P. 1993. Automating the Acquisition of Bilingual Terminology. In *Proceedings of the EACL,* pages 113-119.

Escudero, G., L. Marquez, and G. Rigau. 2000. Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited. In *Proceedings of the 14th European Conference on Artificial Intelligence, ECAI'2000,* pages 421-425.

Fellbaum, C. 1997. Analysis of a hand-tagging task. In *Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Washington, D.C., USA, 1997.

Fillmore, C.J. 1975. An alternative to checklist theories of meaning. In *Papers from the First Annual Meeting of the Berkeley Linguistic Society,* pages 123-132.

Flournoy, R., R. Ginstrom, K. Imai, S. Kaufmann, G. Kikui, S. Peters, H. Schütze, and Y.Takayama. 1998. Personalization and Users' Semantic Expectations. *ACM SIGIR'98 Workshop on Query Input and User Expectations*, Melbourne, Australia.

Frakes, W.B., and R. Baeza-Yates. 1992. *Information Retrieval: Data Structures & Algorithms.* Prentice Hall, Upper Saddle River, NJ.

Frantzi, K. and S. Ananiadou. 1997. Automatic Term Recognition using Contextual Cues. In *Proceedings of the 3rd DELOS Workshop*, Zurich, Switzerland. [http://citeseer.nj.nec.com/frantzi97automatic.html; accessed 09/28/2000]

Franz, M., J.S. McCarley, and S. Roukos. 1998. Ad hoc and multilingual information retrieval at IBM. In *The Seventh Text REtrieval Conference (TREC-8)*, Washington, DC. [http://trec.nist.gov/pubs/trec7/t7_proceedings.html; accessed 09/28/2000]

Franz, M., J.S. McCarley, and R.T. Ward. 1999. Ad hoc, cross-language and spoken document information retrieval at IBM. In *The Eighth Text REtrieval Conference (TREC-8)*, Washington, DC. [http://trec.nist.gov/pubs/trec7/t7_proceedings.html; accessed 09/28/2000]

Fung, P. 1995. Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. In *Proceedings of the Third Workshop on Very Large Corpora*. Boston, MA. [http://www.ee.ust.hk/~pascale/wvlc95.ps; accessed 09/28/2000]

Fung, P. 1998. A statistical view of bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In David Farwell, Laurie Gerber, and Eduard Hovy (ed.), *Third Conference of the Association for Machine Translation in the Americas*, Springer Verlag, pages 1-16. . [http://www.ee.ust.hk/~pascale/amta98.ps; accessed 09/28/2000]

Fung, P., and L.Y. Lee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of COLING/ACL '98*, pages 414-420.

Fung, P., and K. McKeown. 1994. Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping. In *AMTA 94: Partnerships in Translation Technology*, Columbia, Maryland, pages 81-88.

Fung, P., and K. McKeown. 1997. Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*,Hong Kong, pages 192-202.

Furuse, O., and H. Iida. 1992. Cooperation between Transfer and Analysis in Example-Based Framework. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-1992)*, pages 645-651.

Gale, W. and K. Church. 1991. Identifying word correspondence in parallel text. *Proceedings of the DARPA NLP Workshop*. [http://cm.bell-labs.com/cm/ms/departments/sia/doc/93.6.ps; accessed 09/28/2000]

Gale, W.A., K. Church, and D. Yarowsky. 1992a. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26:415-439.

Gale, W.A., K. Church, and D. Yarowsky. 1992b. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92),* pages 101-112.

Gale, W.A., K. Church, and D. Yarowsky. 1992c. Work on statistical methods for word sense disambiguation. In Goldman, R., P. Norvig, E. Charniak, and B. Gale (eds.): *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, AAAI Press, Menlo Park, CA, pages 54-60.

Gale, W.A., and K. W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19(1):75-102.

Gaussier, E. 1998. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In *Proceedings of COLING-ACL-98*, Montreal, pages 444-450. [http://kb.rxrc.xerox.com/publis/mltt/gaussier-netflowcolacl-98.ps; accessed 09/28/2000]

Geeraerts, D. 1993. Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics* 4:223-272.

Gersenovic, M. 1995. The ICD family of classifications. *Methods Inf Med* 34(1/2):172-175.

Gey, F.C., and H. Jiang. English-German cross-language retrieval for the girt collection – exploiting a multilingual thesaurus. In *The Eighth Text Retrieval Conference (TREC-8)*, Washington, DC.

Goldstein, J., and J.G. Carbonell. 1998. The use of mmr and diversity-based reranking in document reranking and summarization. In *Proceedings of the 14th Twente Workshop on Language Technology in Multimedia Information Retrieval*, Enschede, the Netherlands, pages 152-166.

Goldstein, J., V. Mittal, J.G. Carbonell, and M. Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *ANLP/NAACL 2000 Workshop*, pages 40-48.

Golub, G.H., and C.F. van Loan. 1989. *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD.

Graff, D., and R. Finch. 1994. Multilingual Text Resources at the Linguistic Data Consortium. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*. Morgan Kaufmann.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press.

Grefenstette, G. 1996. Light Parsing as Finite-State Filtering. In Wahlster, W. (ed.): *Workshop on Extended Finite State Models of Language*. ECAI-96, Budapest, Hungary.

Gvenir, H.A., and A. Tun. 1996. Corpus-based learning of generalized parse tree rules for translation. In Gordon McCalla, editor, *Proceedings of the Eleventh Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*. volume 1081 of LNAI, Springer Verlag, Berlin, Germany, pages 121-132. [ftp://ftp.cs.bilkent.edu.tr/pub/tech-reports/1996/BU-CEIS-9607.ps.z; unconfirmed]

Gvenir, H.A., and I. Cicekli. 1998. Learning Translation Templates from Examples. *Information Systems*, 23(6):353-363.

Hahn, U., S. Schulz, and M. Romacker. 1999. Part-whole reasoning: A case study in medical ontology engineering. *IEEE Intelligent Systems* 14(5): 59-67.

Hanks, P. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1(1):75-98.

Hanks, P. 2000. Do word meanings exist? *Computers and the Humanities* 34(1/2):205-215.

Harman, D. K. 1993. Overview of the First Text REtrieval Conference (TREC-1). In *The First Text REtrieval Conference (TREC-1)*, NIST Special Publication 500-207.

Harman, D. 1995. The TREC Conferences. In *Proceedings of HIM '95*. Reprint in Sparck-Jones, K., and Willett, P. (eds.): *Readings in Information Retrieval*. Morgan Kaufmann Publishers.

Harman, D. 1997. Overview of the Fifth Text REtrieval Conference (TREC-5). In *The Fifth Text REtrieval Conference (TREC-5)*, NIST Special Publication 500-238.

Hawkins, P., and D. Nettleton. 2000. Large scale WSD using learning applied to SENSEVAL. *Computers and the Humanities* 34(1/2):135-140.

Hearst, M. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Congress on Computational Linguistics*, Nantes, France, pages 539-545.

Hearst, M.A. 1991. Noun homograph disambiguation using local context in large corpora. In *Proceedings of the SEventh Annual Conference of the Centre for the New OED and Text Research: Using Corpora*, Oxford, UK, pages 1-22

Hearst, M.A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23:33-64.

Hearst, M.A., and C. Plaunt. 1993. Subtopic structuring for full-length document access. In *Proceedings of ACM SIGIR 16*, pages 59-68.

Heid, U. *et al*. 1996. Term extraction with standard tools for corpus exploration: Experience from German. In *Proceedings of the 4th International Congress on Terminology and Knowledge Engineering (TKE '96)*, Wien.

Heyn, M. 1996. Integrating machine translation into translation memory systems. In *European Association for Machine Translation - Workshop Proceedings*, ISSCO, Geneva, Switzerland, pages 111-123.

Hirst, G. 1988. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press.

Hovy, E., and C.-Y. Lin. 1997. Automated text summarization in SUMMARIST. ACL (1997), pages 10-17.

Hovy, E., N. Ide, R. Frederking, J. Mariani, and A. Zampolli. 1999. Multilingual Information Management: Current Levels and Future Abilities. [http://www.cs.cmu.edu/~ref/mlim/index.html; accessed 09/28/2000]

Hull, D. 1996. Stemming Algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science* 47(1).

Hull, D. 1997. Automating the Construction of Bilingual Terminology Lexicons. *Terminology* 4(2), 225--244, 1997.

Hull, D. 1999. The TREC-7 Filtering Track: Description and Analysis. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, NIST Special Publication 500-242.

Hull, D.A., and G. Grefenstette. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *19th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 49-57.

IAI Working Paper No. 36: Hybrid Approaches to Machine Translation. Oliver Streiter, Michael Carl, and Johann Haller (ed.) *Institut der Gesellschaft zur Frderung der Angewandeten Informationsforschung e.V. Universitt des Saarlandes*. April 2000. [http://iaisun.iai.uni-sb.de/~carl/iaiwp/index.html and http://rockey.iis.sinica.edu.tw/oliver/iaiwp/; accessed 09/28/2000]

Ide, N. 1999. Parallel Translations as Sense Discriminators. In *Proceedings of SIGLEX99*, Washington D.C, USA

Ide, N. 2000. Cross-lingual Sense Determination: Can it work? *Computers and the Humanities* 34(1/2): 223-234.

Iomdin, L. and O. Streiter. 1999. Learning from Parallel Corpora: Experiments in Machine Translation. In *Proceedings of the Dialogue'99 International Seminar in Computational Linguistics and Applications*, Tarusa (Russia). [http://proling.iitp.ru/bibitems and http://rockey.iis.sinica.edu.tw/oliver/publ; accessed 09/28/2000]

Jacquemin, C. 1996. What is the TREE that we see through the Window: A Linguistic Approach to Windowing and Term Variation. *Information Processing and Management* 32(4):445-458.

Jorgenson, J.C. 1998. The Psychological Reality of Word Senses. *Journal of Psycholinguistic Research* 19(3):167-190.

Jung, H.-M., S. Yuh, C.-M. Sim, T. Kim, and D.-I. Park. 1998. A domain identifier using domain keywords from balanced web documents. In *First International Conference on Language Resources & Evaluations*. Granada, Spain.

Justeson, J. and S. Katz. 1995. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* 1(1), pages 9-27.

Kando, N., and T. Nozue (ed.) 1999. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*. Nihon Printing Co., Ltd., TOKYO, Japan. [http://www.rd.nacsis.ac.jp/~ntcadm/workshop/OnlineProceedings/; accessed 09/28/2000]

Kando, N., K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, S. Hidaka, and J. Adachi. 1999. The NTCIR Workshop: The First Evaluation Workshop on Japanese Text Retrieval and Cross-Language Information Retrieval. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*.

Karov, Y., and S. Edelman. 1996. Learning similarity-based word disambiguation from sparse data. In *Proceedings of the Fourth Workshop on Very Large Corpora.*

Karov, Y., and S. Edelman. 1998. Similarity-based word sense disambiguation. *Computational Linguistics* 24(1):41-59.

Kaufmann, S. 2000. Second Order Coherence. *Computational Intelligence* 16(4):511-524.

Kilgariff, A. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language* 12(4), Special Issue on Evaluation.

Kilgariff, A., and J. Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities* 34(1/2): 15-48.

Kilgariff, A., and M. Palmer. 2000. Introduction to the special issue on SENSEVAL. *Computers and the Humanities* 34(1/2):1-13.

Kim and Moldovan. 1993. Acquisition of Semantic Patterns for Information Extraction from Corpora. In *Proceedings of the 9th IEEE Conference on AI for Applications*. IEEE Computer Society Press.

Kinoshita, S., A. Kumano, and H. Hideki. 1994. Improvement in Customizability using Translation Templates. In *Proceedings of COLING'94*. Kyoto, Japan. pages 25-30.

Kirsten, W. and R. Klar. 1996. *Documentation und Informationsaufbereitung für den Arzt*. Epsilon Verlag, Darmstadt.

Kitano, H. 1993. A Comprehensive and Practical Model of Memory-Based Machine Translation. In Ruzena Bajcsy (Ed.) *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann V.2 (1993) 1276-1282.

Klavans, J.L., and J. Shaw. 1995 Lexical semantics in summarization. *In Proceedings of the First Annual Workshop of the IFIP Working Group FOR NLP and KR*, Nantes, France.

Klein and ... 2000. Extracting Drug-ADR Relations, a Semantic Approach.

Kozima, H. 1994. *Computing Lexical Cohesion as a Tool for Text Analysis.* Phd Thesis, University of Electro-Communications.

Kozima, H., and T. Furugori. 1993. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of EACL 6*, Utrecht, The Netherlands, pages 232-239.

Kumano, A., and H. Hirakawa. 1994. Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information. In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan.

Kupiec, J. 1993. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proceedings of the ACL*, pages 17-22.

Kupiec, J.M., J. Pedersen, and F. Chen. 1995. A trainable document summarizer. *In Proceedings of the 18th Annual Int. ACM/SIGIR Conference on Research and Development in IR*, Seattle, WA, pages 68-73.

Lancaster, W. F. 1969. MEDLARS: Report on the evaluation of its operating efficiency. *American Documentation* 20, pages 119-142.

Landauer, T.K., and M.L. Littman. 1990. Fully Automatic Cross-Language Document Retrieval using Latent Semantic Indexing. In *Proceedings of the Sixth Annual Conference of the Centre for the New Oxford English Dictionary and Text Research*.

Langkilde, I., and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of COLING-ACL'98*.

Lauer, M. 1995. Corpus Statistics Meet the Noun Compound: Some Empirical Results. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 47-54.

Lauriston, A. 1994. Automatic Recognition of Complex Terms: Problems and the TERMINO Solution. *Terminology* 1(1), pages 147-170.

Leacock, C., G. Towell, and E.M. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Techonology*, Morgan Kaufman: San Francisco, CA.

Leacock, C., G. Towell, and E.M. Voorhees. 1996. Towards building contextual representations of word senses using statistical models. In Boguraev, B., and J. Pustejovsky (eds.), *Corpus Processing for Lexical Acquisition.* MIT Press: Cambridge, MA.

Leacock, C., M. Chodorow, and G.A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics* 24(1): 147-165.

Lee, L. 1997. Similarity-Based Approaches to Natural Language Processing. PhD thesis, Harvard University. (also Technical Report TR-11-97). [ftp://das-ftp.harvard.edu/techreports/tr-11-97.ps.gz; accessed 09/28/2000]

Lee, L. 1999. Measures of distributional similarity. In *Proceedings of ACL 37*, pages 25-32.

Lee, L., and F. Pereira. 1999. Distributional similarity: Clustering *vs.* nearest neighbors. In *Proceedings of ACL 37*, pages 33-40.

Lehnert, W. 1991. Symbolic/Subsymbolic Sentence Analysis: Exploiting the Best of Two Worlds. In J. Barnden and J. Pollack (eds.): *Advances in Connectionist and Neural Computation Theory*. Vol. 1., Ablex Publishers, Norwood, NJ.

Lesk, M.E. 1986. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cone. In *Proceedings of the SIGDOC Conference*.

Lin, D. 1998. Automatic Retrieval and Clustering of Similar Words. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL '98), pages 768-774.

Lindberg, D.A., B.L. Humphreys, and A.T. McCray. 1993. The unified medical language system. *Methods Inf Med* 32(4):281-291.

Linguistic Data Consortium. 1997. Hansard Corpus of Parallel English and French. [http://www.ldc.upenn.edu/; accessed 09/28/2000]

Luhn, P.H. 1958 Automatic creation of literature abstracts. *IBM Journal*, pages 159-165.

Luz, C. 1997. *Vom Freitext zu Kode.* Epsilon Verlag, Darmstadt.

Macklovitch, E. 1994. Using Bi-textual Alignment for Translation Validation: The TransCheck System. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, Columbia, MD.

Maedche, A., and S. Staab. 1999. Discovering Conceptual Relations from Text. In:

Mani, I., and E. Bloedern. 1997. Multi-document summarization by graph search and merging. In *Proceedings of AAAI-97*, pages 622-628.

Manning, C.D., and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA.

Marcu, D. 1997. From discourse structures to text summaries. ACL (1997), pages 82-88.

Marcus, M., M., B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn-Treebank. *Computational Linguistics* 19(2).

Maruyama, H., and H. Watanabe. 1992. Tree Cover Search Algorithm for Example-Based Translation. In Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT. Montreal, pages 173-184.

McKeown, K.R., J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. 1999. Towards Multidocument Summarization by Reformulation: Progress and Prospects. In *Proceedings of AAAI-99*, Orlando, FL, pages 453-460.

McKeown, K.R., J. Robin, and K. Kukich. 1995. Designing and evaluating a new revision-based model for summary generation. *Info. Proc. and Management*, 31(5).

McLean, I.J. 1992. Example-Based Machine Translation using Connectionist Matching. In *TMI-92*.

Melamed, I.D. 1996. A Geometric Approach to Mapping Bitext Correspondences. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing (NeMLaP)*. Philadelphia, PA.

Melamed, I.D. 1997. A Word-to-Word Model of Translational Equivalence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL'97)*. Madrid. pages 490-497. [ftp://ftp.cis.upenn.edu/pub/melamed/papers/transmod.ps.gz; accessed 09/28/2000]

Melamed, I.D., and P. Resnik. 2000. Tagger evaluation given hierarchical tag sets. *Computers and the Humanities* 34(1/2):79-84.

Messing-Junger, A.M. 1998. Procedure Coding System: Background and development. *Langenbecks Arch Chir Suppl Kongressbd* 115:757-763.

Miller, G.A. WordNet: A Lexical Database for English. *Communications of the ACM* 11.

Mitra, M., A. Singhal, and C. Buckley. 1997. Automatic text summarization by paragraph extraction. ACL (1997).

Mizzaro,S. (1998): How many relevances in information retrieval? *Interacting With Computers* 10(3).

Moon, R. 2000. Lexicography and disambiguation: The size of the problem. *Computers and the Humanities* 34(1/2):99-102.

Mooney, R.J. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, pages 82-91.

Moore, W., U.N. Riede, R.A. Polacsek, R.E. Miller, and G.H. Hutchins. 1992. *Automated Translation of German to English Medical Text.* Baltimore, MD.

Morin, E. 1999. Using Lexico-syntactic Patterns to Extract Semantic Relations between Terms from a Technical Corpus. In *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE '99)*, Innsbruck, Austria, pages 268-278.

Morris, J., and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indication of the structure of text. *Computational Linguistics* 17(1):21-48.

Murata, M., Q. Ma, K. Uchimoto, and H. Isahara . 1999. An Example-Based Approach to Japanese-to-English Translation of Tense, Aspect, and Modality. In *TMI'99*.

Nagao, M.A. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In Elithorn, A., and Bernerji, R., eds., *Artificial and Human Intelligence*. North-Holland.

Nagao, M.A. 1985. Framework of a Mechanical Translation between Japanese and English by Analogy Principle.

Nakagawa, H. and T. Mori. 1998. Nested Collocation and Compound Noun for Term Extraction. In *Proceedings of the First Workshop on Computational Terminology, ACL-COLING '98*, pages 64-70.

Nédellec, C. 1999. Corpus-based learning of semantic relations by the ILP system Asium. In *Proceedings of the Learning Language in Logic workshop, ICML '99*, pages 28-39. [http://www.lri.fr/Francais/Recherche/ia; accessed 09/28/2000]

Ng, H.T. 1997. Getting serious about word sense disambiguation. In *Proceedings of the ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Washington D.C., USA.

Ng, H.T. 1999. A case study on inter-annotator agreement for word sense disambiguation. Washington D.C., USA, 1999.

Ng, H.T., and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of ACL96*.

Nie, J.-Y., M. Simard, P. Isabelle, and R. Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. *In The Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, pages 74-75.

Nirenburg S. et al. 1993. Two Approaches to Matching in Example-Based Machine Translation. *Proceedings of TMI-93*, Kyoto, Japan.

Nirenburg, S., S. Beale, and C. Domashnev. 1994. A Full-Text Experiment in Example-Based Machine Translation. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, England, pages 78-87. [http://clr.nmsu.edu/users/sb/papers/ebmt/col94/col94.ps; accessed 09/28/2000]

Niwa, Y., and Y. Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of COLING 94*, pages 304-309.

Oard, D. 1997. Cross-Language Text Retrieval Research in the USA. [http://www.enee.umd.edu//medlab/filter/papers/delos/paper.html; accessed 09/28/2000]

Oard, D. 1998. A comparative study of query and document translation for cross-language information retrieval. In *The Third Conference of the Association for Machine Translation in the Americas (AMTA)*, Philadelphia, PA.

Over, P. 1997. TREC-5 Interactive Track Report. In *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, NIST Special Publication 500-238.

Oueslati. 1996. Terminology and semantic relation extraction from texts In: *KAW96*

Paice, C.D. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Info. Proc. and Management*, 26:171-186.

Pedersen, T., and R. Bruce. 1997. A new supervised learning algorithm for word sense disambiguation. In *Proceedings of AAAI*.

Peters, C., and E. Picchi. 1995. Capturing the Comparable: A System for Querying Comparable Text Corpora. In *Proceedings of Analisi Statistica dei Dati Testuali (JADT)*, pages 247-254.

Peters, W., I. Peters, and P. Vossen. 1998. Automatic Sense clustering in EuroWordNet. In *Proceedings of the First Interational Conference on Language Resources and Evaluation*, Granada.

Qiu, Y. and H.P. Frei. 1993. Concept Based Query Expansion. In *Proceedings of Sixteenth Annual ACM SIGIR Conference*, Pittsburgh, PA, pages 160-169.

Radev, D.R., and K.R. McKeown. 1998. Generating natural language summaries from multiple online sources. *Computational Linguistics*, 24(3):469-501.

Radev, D.R., H. Jing, and M. Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *ANLP/NAACL 2000 Workshop*, pages 21-29.

Rapp, R. 1995. Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the ACL35*, Boston, Mass, pages 321-322.

Rasmussen, E. 1992. Clustering Algorithms. In Frakes and Bayeza-Yates (1992).

Rayner, M. and P. Bouillon. 1995. Hybrid Transfer in an English-French Spoken Language Translator. In *Proceedings of IA'95*, Montpellier, France. [http://www.cam.sri.com/; accessed 09/28/2000]

Reinke, U. 1999. Evaluierung der linguistischen Leistungsfhigkeit von Translation Memory-Systemen -- Ein Erfahrungsbericht. In *LDV-Forum, Forum der Gesellschaft fr Linguistische Datenverarbeitung (GLDV)*, 1-2 (in German) [http://rockey.iis.sinica.edu.tw/oliver/iaiwp/p14/reinke1.html; accessed 09/28/2000]

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*.

Resnik, P. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Washington, D.C., USA.

Resnik, P. 1998. Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text. [ftp://ftp.cs.umd.edu/pub/papers/papers/ncstrl.umcp/CS-TR-3922/CS-TR-3922.ps.Z; accessed 09/28/2000]

Resnik, P. 1999. Mining the web for bilingual text. In 37th Annual Meeting of the Association for Computational Linguistics (ACL'99).

Resnik, P., and D. Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Washington, D.C. 1997.

Resnik, P., and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering* 5(2).

Reynar, J.C. 1998. Topic Segmentation: Algorithms and Applications. Phd Thesis, University of Pennsylvania.

Riloff, E. 1993. Automatically Constructing a Dictionary for Information Extraction Tasks. In *Proceedings of the 11th National Conference on AI (AAAI-93).*

Riloff, E. 1996. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the 13th National Conference on AI (AAAI-96).*

Riloff, E., and R. Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the 16th National Conference on AI (AAAI-99).*

Rindfleisch. 2000. Extracting Molecular Binding Relationships from Biomedical Text

Ruch, P., J. Wagner, P. Bouillon, R.H. Baud, A.M. Rassinoux, and J.R. Scherrer. 1999. MEDTAG: Tag-like semantics for medical document indexing. In *Proc AMIA Symp*, pages 137-141.

Sacaleanu, B. Forthcoming. *Domain specific tuning and extension of lexical semantic resources*. MA Thesis, University of the Saarland.

Salton, G. 1970. Automatic Processing of Foreign Language Documents. Journal of American Society for Information Sciences, 21:187-194.

Salton, G., A. Singhal, C. Buckley, and M. Mitra. 1996. Automatic text decomposition using text segments and text themes. In *Proceedings of the 7th ACM Conference on Hypertext*, 53-65.

Sato, S. 1991a. Example-Based Machine Translation. Ph.D. Thesis, Kyoto University.

Sato, S. 1991b. Example-Based Translation Approach. Proceedings of FGNLP-91, ATR Interpreting Telephony Research Laboratories, pages 1-16.

Sato, S. 1993. Example-Based Translation of Technical Terms. Proceedings of TMI-93, Kyoto, pages 58-68.

Sato, S., and M. Nagao. 1990. Towards memory-based translation. In *COLING-90*, Helsinki, Finland, vol. 3, pages 247-252.

Schäuble, P., and D. Knaus. The Various Roles of Information Structures. In *Information and Classification*, pages 282-290.

Schäuble, P. and P. Sheridan. 1998. Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the Sixth Text Retrieval Conference (TREC6)*. NIST Special Publication 500-240.

Schütze, H. 1992. Context space. In Goldman, R., P. Norvig, E. Charniak, and B. Gale (eds.): *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, AAAI Press, Menlo Park, CA, pages 113-120.

Schütze, H. 1997. Ambiguity Resolution in Language Learning: Computational and Cognitive Models. Stanford: CSLI Publications.

Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97-123.

Schütze, H. and J. Pedersen. 1997. A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval. *Information Processing and Management* 33(3), pages 307-317.

Sekine, S., and Isahara, H. 2000. IREX: IR and IE Evaluation Project in Japanese. In Proceedings of the 2nd International Conference on Language Resources & Evaluation (LREC 2000).

Shaw, J. 1995. Conciseness through aggregation in text generation. In *Proceedings of 33rd Association for Computational Linguistics*, pages 329-331.

Sheridan, P., and J.P. Ballerini. 1996. Experiments in multilingual information retrieval using the spider system. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, The Association for Computing Machinery, New York, NY, pages 58-65.

Sheridan, P., M. Wechsler, and P. Schauble. 1997. Cross-language speech retrieval: establishing a baseline performance. In *20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 99-107.

Sheridan, P., J.P. Ballerini, and P. Schäuble. 1998. Building a Large Multilingual Test Collection from Comparable News Douments. In G. Grefenstette, (ed.): *Cross-Language Information Retrieval*, Chapter 11. Kluwer Academic Publishers, Boston, MA.

Smadja, F. 1992. How to Compile a Bilingual Collocational Lexicon Automatically. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*.

Smadja, F., *et al.* 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1):1-38.

Small, S.L. 1980. *Word Expert Parsing: A Theory of Distributed Word-based Natural Language Understanding.* Ph.D. thesis, The University of Maryland, Baltimore, MD.

Small, S.L. 1983. Parsing as cooperative distributed inference. In King, M. (ed.): *Parsing Natural Language.* Academic Press, London.

Soderland, S. 1999. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning: Special Issue on Natural Language Processing* 34(1/3), 233-272.

Soderland, S., D. Fisher, J. Aseltine, and W. Lehnert. 1995. CRYSTAL: Inducing a Conceptual Dictionary. In *Proceedings of the 14th International Joint Conference on AI (IJCAI-95).*

Sparck Jones, K. and C. van Rijsbergen. 1975. Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection. In *British Library Research and Development Report 5266,* Computer Laboratory, University of Cambridge.

Stein, G.C., T. Strzalkowsi, and G.B. Wise. 1999. Summarizing multiple documents using text extraction and interactive clustering. In *Proceedings of the Conference Pacific Association for Computation Linguistics (PACLING),* pages 200-1208..

Streiter, O., L. L. Iomdin, M. Hong, and U. Hauck. 1999. Learning, Forgetting and Remembering: Statistical Support for Rule-Based MT. In *Proceedings of TMI'99.* Chester, England. 44-54. [http://rockey.iis.sinica.edu.tw/oliver/publ and http://proling.iitp.ru/bibitems; accessed 09/28/2000]

Streiter, O., M. Carl, L. L. Iomdin. 2000. A Virtual Translation Machine for Hybrid Machine Translation. In *Proceedings of the Dialogue'2000 International Seminar in Computational Linguistics and Applications*. Tarusa, Russia. [http://rockey.iis.sinica.edu.tw/oliver/publ and http://proling.iitp.ru/bibitems; accessed 09/28/2000]

Strzalkowski, T., J. Wang, and G.B. Wise. 1998. A robust practical text summarization system. In *AAAI Intelligent Text Summarization Workshop,* Stanford, CA, pages 26-30.

Sumita, E., and H. Iida. 1991. Experiments and Prospects of Example-Based Machine Translation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. pages 185-192.

Swanson, D.R., and N.R. Smalheiser. 1997. An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. *Artificial Intelligence* 91.

Tait, J.I. 1983. *Automatic Summarizing of English Texts*. PhD thesis, University of Cambridge, Cambridge, UK.

Tanaka, K. 1996. Extraction of Lexical Translations from Non-Aligned Corpora. In *Proceedings of COLING-96.* [http://www.etl.go.jp/~kumiko/Publications/coling96.ps.gz; accessed 09/28/2000]

TDT. 1997a. The TDT Pilot Study Corpus Documentation Version 1.3. Distributed by the Linguistic Data Consortium.

TDT. 1997b. The Topic Detection and Tracking (TDT) Pilot Study Evaluation Plan.

Teufel, S., and M. Moens. 1997. Sentence extraction as a classification task. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, pages 58-65.

Thier, A. 1999. Medias Web Service. Medizinisches Informations- und Auskunfts-System. Ein Arzt- und Behandlungsspezifischer Web-Agent zur Automatischen Informationsrecherche und -Aufbereitung im Internet/Intranet. Diplomarbeit, Technische Universität Darmstadt.

Tiedemann, J. 1998a. Extracting Phrasal Terms using Bitext. [http://stp.ling.uu.se/~corpora/plug/paper/WTRC.ps.gz; accessed 09/28/2000]

Tiedemann, J. 1998b. Extraction of Translation Equivalents from Parallel Corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics (NODALI98)*. pages 120-128. [http://stp.ling.uu.se/~joerg/paper/Nodalida98.ps.gz; accessed 09/28/2000]

TIPSTER: [http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/; accessed 09/28/2000]

Tomuro, N. 2000. Automatic extraction of systematic polysemy using tree-cut. In *Proceedings of the workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems at Language Technology Joint Conference, Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL2000)*, Seattle, WA, pages 20-27.

Turkato, D. 2000. Automatically Creating Bilingual Lexicons for Machine Translation from Bilingual Text.

Utsuro, T., H. Ikeda, M. Yamane, Y. Matsumoto, and M. Nagao. 1994. Bilingual text matching using bilingual dictionary and statistics. In *15th COLING*, pages 1076-1082. [http://cactus.aist-nara.ac.jp/lab/papers/utsuro/bitext9408.ps.gz; accessed 09/28/2000]

Veale, T., and A. Way. 1997. Gaijin: A Bootstrapping Approach to Example-Based Machine Translation. In *Proceedings of the NeMNLP'97, New Methods in Natural Language Processing*. Sofia, Bulgaria. [http://www.compapp.dcu.ie/~tonyv/papers/gaijin.html; accessed 09/28/2000]

Véronis, J. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop, Herstmonceux Castle (England)*, pages 2-4.

Voorhees, E. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Voorhees, E., and D. Harman. 2000. Overview of the Eighth Text REtrieval Conference (TREC-8), In *Proceedings of the Eighth Text Retrieval Conference (TREC8)*. To appear.

Vossen, P. (ed.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers. Reprinted from *Computers and the Humanities* 32:2-3.

Wiebe, J., R. Bruce, and L. Duan. 1997. Probabilistic event categorization. In *Proceedings of the Second International Conference on Recent Advances in NLP (RANLP 97*), Tzigov Chark, Bulgaria.

Wiebe, J., J. Maples, L. Duan, and R. Bruce. 1997. Experience in WordNet sense tagging in the Wall Street Journal. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?* Washington, D.C., U.S.A.

Wilks, Y., D. Fass, C.M. Guo, J.E. McDonald, T. Plate, and B.M. Slator. 1990. Providing machine tractable dictionary tools. *Machine Translation* 5(2):99-154.

Wilks, Y., B. Slator, and L. Guthrie. 1996. *Electric Words: dictionaries, computers and meanings.* Cambridge, MA: MIT Press.

Wu, D. 1995. Grammarless extraction of phrasal translation examples from parallel texts. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation.* Leuven, Belgium. [ftp://ftp.cs.ust.hk/pub/dekai/tmi95.Wu.ps.Z; accessed 09/28/2000]

Xu, J. and W. Bruce Croft. 2000. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems* 18(1), pages 79-112.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1(1/2):67-88.

Yang, Y., and C.G. Chute. 1993. Words or concepts: The features of indexing units and their optimal use in information retrieval. In *Proceedings of SCAMC'93*, pages 685-689.

Yang, Y., and C.G. Chute. 1994a. An example-based mapping method for text categorization and retrieval. *ACM Transaction on Information Systems (TOIS)*, 12(3):252-277.

Yang, Y., and C.G. Chute. 1994b. Words or concepts: The features of indexing units and their optimal use in information retrieval. In *Journal of AMIA 1994 (SCAMC'94)*, 18(Symp.Suppl):157-161.

Yang, Y., J.G. Carbonell, R.D. Brown and R.E. Frederking. 1998. Translingual Information Retrieval: Learning from Bilingual Corpora. In *Artificial Intelligence Journal*, 103:323-345. [http://www.cs.cmu.edu/~ralf/papers/aij98.ps; accessed 09/28/2000]

Yang, Y., R.D. Brown, R.E. Frederking, J.G. Carbonell, Y. Geng and D. Lee. 1997. Bilingual-corpus Based Approaches to Translingual Information Retrieval. In *Proceedings of MULSAIC'97.*

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget's categories. In *Proceedings of COLING-92*, Nantes, France.

Yarowsky, D. 1993. One sense per collocation. In *Proceedings of the ARPA Workshop on Human Language Techonology,* Morgan Kaufman: San Francisco, CA.

Yarowsky, D. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, pages 88-95.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings ACL 33*, Cambridge, MA.

Yeh, A. 2000. ComparingTwo Trainable Grammatical Relations Finders. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, Saarbrücken, Germany.

Zavrel, J., and W. Daelemans. 1997. Memory-based learning: Using similarity for smoothing. In *Proceedings of ACL 35*, pages 436-443.

Zeng, Q. and J.J. Cimino. 1998. Automated knowledge extraction from the UMLS. In *Proc AMIA Symp* pages 568-572.

Zeres GmbH, Bochum, Germany 1997. ZERESTRANS Benutzerhandbuch.

Zhai, C. 1997. Fast Statistical Parsing of Noun Phrases for Document Indexing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington DC*, pages 312-319.

Zhai, C. *et al*. 1997. Evaluation of Syntactic Phrase Indexing - CLARIT NLP Track Report. In *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pages 347-357.