

# Automatic Multilingual Indexing and Classification

Bärbel Ripplinger & Dieter Maas

IAI

Martin-Luther-Str. 14

D-66111 Saarbrücken

Germany

email:{babs|dieter}@iai.uni-sb.de

## Abstract

Most of today's published scientific and technical articles are written in English. Therefore, the number of English documents being collected by information brokers such as bibliographic database producers, libraries and publishers increases rapidly. However, there will still be a number of documents only available in the native language of the author. One method to facilitate access to this information with reliable recall and precision is provided by smart indexing and classification processes. The system AUTINDEX offers various tools for monolingual as well as for multilingual indexing and classification by taking advantage of sophisticated language processing technologies and already existing special purpose language resources such as thesauri, classification schemes and large lexicons.

**Keywords: Indexing, Classification, Linguistic Processing, Multilinguality**

## 1. Objectives

Most of today's published scientific and technical articles are written in English. Therefore, the number of English documents being collected and maintained by information brokers/providers such as bibliographic database producers, providers (libraries) and publishers increases rapidly. However, there is always a number of documents only available in the native language of an author. In addition, the competing, often free-of-charge internet information flood, presses information producers to offer better services and to broaden the customer base of their online products. To guarantee a high quality service through advanced search facilities with intuitive communication interfaces that enable users to manage their information needs efficiently and easily, the information providers have to look for time and cost effective automation methods and procedures.

One method to facilitate the access to the information with a reliable recall and precision is provided by smart indexing processes. This will not only ensure a consistent indexing in one language but in multiple languages and thus allows for the multilingual presentation of the information. Using the same repository of terms for all languages preferably contained in a multilingual thesaurus enables a fast and reliable cross-language retrieval of the information, i.e. the user will get documents not only in the query language but all relevant documents in any language.

At present no commercial system supports the human indexer in his/her intellectual task, so most publications (journals, conference papers, dissertations, reports, and books) are indexed intellectually by assigning descriptors in the respective language. This manual intellectual indexing is time consuming, expensive, and inhomogeneous because of the different background knowledge and expertise of the personnel.

The AUTINDEX system, developed at the IAI, supports information producers by using a generic solution to automatically index and classify documents in different languages. This system allows the quicker, cheaper, and consistent population of information repositories. To do this, AUTINDEX takes advantage of sophisticated language processing technologies and already existing special purpose language resources such as thesauri, classification schemes and large lexicons. Due to a modular software design, the system provides utilities for monolingual indexing and classification and for a parallel multilingual indexing and classification system, currently only for English and German. The consistency of the multilingual indexing will be guaranteed by using the same resources.

Using such a utility the return on investment will be enormous taking into account that such a system can automate the task of indexing and classifying a document. Therefore a system for automatic indexing and classification saves time and money, i.e. the proposed system will help people achieve more in less time, and thus increases human productivity. Time saving is an advantage for the actuality of the database information. Money saving is an advantage to improve information completeness, especially of the German-language literature, now searchable also in English or even in other European languages.

With the growing flood of information it becomes more and more important to find the right information. One major solution to this problem can be provided by applying language technologies and, in this case, smart indexing. Through an incorporation of language technologies, as proposed here, not only is the human indexer supported in his task, but the impact for information consumers is considerable. The better the documents are indexed, the more precise the search results become and the greater the acceptance by the customers. Hence, every service dealing with information dissemination, for instance search engines like Yahoo, Lycos or Altavista, publishers, bibliographic database producers and libraries, e.g. industry and public services in the document management and workflow domain could take advantage of such a system that supports multilingual automatic indexing. Currently indexing is done in most cases manually and separately for each language [9]. Using the existing machine-readable thesauri and classification scheme also for the automatic indexing and classification, there will be no change of the interface but customers will get better and more reliable results. Smaller service providers like news agencies that provide collections of documents for their users according to profiles are also potential clients. Here, not only will the underlying technique be improved but also the profiling can be done by using the descriptors of the thesaurus as representing the areas of interest. Additionally, this technique can be used to enhance the presentation aspects, i.e. the user guidance can be improved through a more sophisticated agent and portal technologies. Using the same resource for indexing and profiling will facilitate and speed up the information search due to a better knowledge management. Further, the thesaurus can also be used to give the users a clear indication of what kind of information they can search in the databases and thereby improve the transparency. This, in turn, considerably increases the user-friendliness with regard to both the usage and the results

## 2. Related Work

Commercial indexing systems such as CINDEK or MACREX support the human indexer in the index preparation and the processing (editing and formatting) of manually produced indexes. Most of them provide also a spell-checking facility. But the time consuming intellectual task – the assigning of descriptors to documents – is only supported by maintaining the actual list of terms used for the indexing.

There are some current research activities investigating different approaches to enable automatic indexing and its deployment in information retrieval systems. The most ambitious work is the latent semantic indexing (LSI) approach [3][5]. This method assumes an underlying or *latent* structure in the pattern of word usage across documents. Based on this information, LSI constructs a term-document matrix to represent similarities of contexts in which words appear. Because these word associations are derived from a numerical analysis of the considered documents there is no need to use any external dictionary, thesaurus or knowledge base. To use this approach in a multilingual environment, it is necessary to have a set of parallel documents to compute the basic set of cross-language associations which means a major drawback for a real life application of this approach.

At the University of California at Berkeley [8], an indexing method based on lexical associations is being developed using a controlled vocabulary. To identify these associations, statistical methods are applied to create a dictionary of associations between lexical items contained in the titles, authors and abstracts and the controlled vocabulary which was extracted for records made by human indexers. This approach has only been proven on monolingual data, and there are limitations reported related to number of topics assigned due to the lack of more sophisticated natural language processing techniques.

Within the Condorcet project [1] carried out at the University of Twente, a so-called *controlled-term approach* is used to index scientific documents. Based on a *structured*

*ontology*, concepts and relations rather than lexical items are used as indexes. This means after the tagging of the document each lexical item has to be mapped to the proper concepts and/or relations. This is done by determining the syntactic structure and the deep structure of a sentence to get information about semantic roles. By means of a knowledge-based module, this deep structure will then be mapped onto index terms. This approach increases the precision because concepts are mostly language-independent and non-ambiguous. Taking also relations between concepts into account for the indexing means that thesauri used for such an approach have to have such relations, which is definitely a deficiency of most classical thesauri available. Only a few of them such as UMLS in the medical field or ARGOVOC can provide such a structure. Another drawback could be the knowledge-based module which has to perform deductive matching, therefore this approach is proven to be only successfully applicable to a particular domain. There is also no work carried out in a multilingual environment.

### 3. The Approach

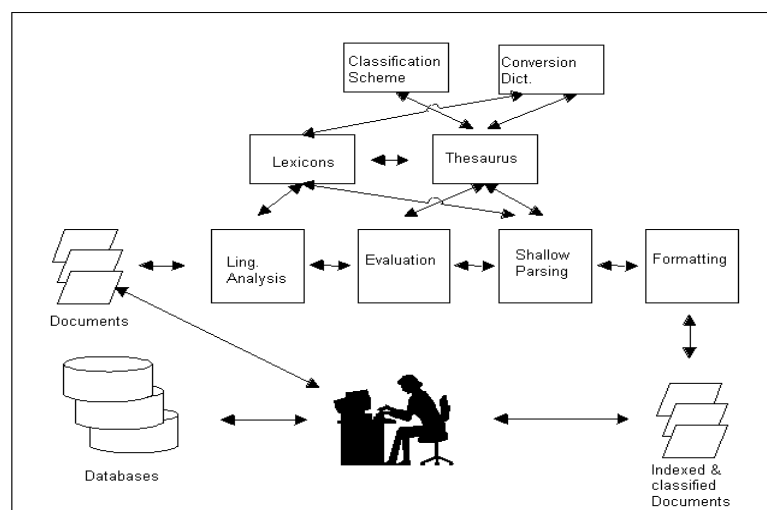
#### 3.1. The Architecture of AUTINDEX System

AUTINDEX takes advantage of sophisticated language processing technologies and already existing special purpose language resources such as thesauri, classification schemes and large lexicons.

The approach is based on a controlled vocabulary and existing advanced natural language processing technologies. The controlled vocabulary is provided by a classical thesaurus together with a specialised multilingual. The linguistic processing provides all the information necessary to assign the concepts of the thesaurus to the words (including multiword units) of the documents. No knowledge base or extensive preprocesses to determine the necessary relationships are needed. The indexing approach is additionally enhanced with a shallow statistical technique. Classification is also based on the output of the linguistic processing and an already existing classification scheme.

Due to a modular software design, the system provides utilities for monolingual indexing and classification and for a parallel bilingual indexing and classification system for English and German. The consistency of the bilingual indexing will be guaranteed by using the same resources.

The illustration below shows how the various components of the AUTINDEX system interact and which resources they use. The role of the human indexer within this system is to select the documents for the indexing and to evaluate the result of the automatic indexing. Indexing is related only to identify keywords (known thesaurus concepts) and free terms (candidates for thesaurus concepts) from the content of a document. The only bibliographic



information used is the title.

Each module of the system, the German indexing and classification, the English, and the bilingual (English to German, and German to English) module underlies the same language technology components as described below, whereas they take the specific characteristics of each language (for instance, multiword units as terms) into account. The bilingual module is based on the two monolingual modules, and will provide the human indexer with the translations of the terms in the particular foreign language.

This architecture is already evaluated and tested with the German component, the English and the bilingual English-German component will be enhanced and tested within the forthcoming project BINDEX, funded by EC (IST-1999-20028).

### 3.2. The Resources

The AUTINDEX system uses the following resources<sup>1</sup>:

➤ Thesauri:

The system provides a special API to integrate different monolingual but also bilingual thesauri, i.e. users can work with their own thesaurus, in their preferred language.

➤ Classification Schemes:

Also the classification schemes can be determined and integrated depending on the user's requirement by means of an API.

➤ Lexical Resources

Monolingual and bilingual lexicons used the system are developed and maintained by the IAI. Each entry contains morphological, syntactic and semantic information used by the components described below. The entries have also special domain marker which are used for translation purposes. Currently the monolingual German lexicon contains 37.000 morphemes, the English 38.000. The coverage of the English-German lexicon consists of 480.000 entries (simple words and multiword units).

### 3.3. The Linguistic Components

AUTINDEX uses the following components to carry out indexing and classification:

The first three components described below also represent the different processes which are sequentially applied resulting in the indexing and classification of a document which can then be further processed by a human indexer.

➤ Linguistic Analysis

The first step the system performs is the linguistic analysis which consists of a lemmatisation, a part of speech tagging and a homograph analysis for German texts. Linguistic information is assigned to each word of the document information such as the word class, semantic features as well as derivational and decomposition information for compound words. This analysis is performed with the *MPRO* tool developed at IAI and available for several languages [6]. During this analysis (within the segmentation phase), the document language is determined.

➤ Evaluation

In the second step, the text is weighted by means of the thesaurus. For each meaningful word (words with word class noun, verb or adjective), the semantic features assigned during the linguistic analysis are collected. Then the most frequent semantic features are computed, and all words of the text, which have such a feature are collected. Functional words are excluded from this process. This set of so-called *keywords* is then evaluated against the thesaurus. Thereby thesaurus-specific hierarchical information, e.g. broader terms, narrower terms and synonym relations, is used to generate a set of descriptors (indexing). Additionally, branch codes, country codes, and so called facets (topics) extracted from the classification scheme are attached to the document (classification).

---

This weighting approach takes not only complete words into account for this statistics but also the parts of compound words which can be extracted from the compositional information provided by the analysis in step 1. The decomposition process is available for all languages, i.e. for English compounds which are mostly multiword-units, a special treatment is developed (cf. below).

➤ Shallow Parsing

In a third step, a term extraction will be carried out to identify multiword units and their respective syntactic variants (*measurement procedure* vs. *procedure of measurement* vs. *procedure to measure*). Thus, the document is parsed. From the result of this process only nominal and prepositional phrases as well as complex nouns are extracted as term candidates. These terms are then evaluated against the thesaurus if their semantic features belong to the relevant class determined in step 2.

The bilingual indexing, i.e. identifying German descriptors of English documents (or vice versa), is done by applying the steps above to English documents and using then the bilingual thesauru, or a monolingual thesaurus together with the bilingual dictionary and the provided domain markers to generate the appropriate descriptors and free terms (cf. above).

### 3.3 Result

The result of the former steps is a structured list as shown below consisting of different sorts of data: the set of descriptors (concept of the thesaurus), free terms, the classification information but also the unknown words together with the document itself:

<b>INPUT</b>	<p><b>Title:</b> A leaky-mode S-parameter extraction technique for efficient design of the microstrip line leaky-wave antenna</p> <p><b>Abstract:</b> A leaky-mode S-parameter extraction technique is proposed in this paper. The proposed method can be employed in a numerical way or an experimental way. By properly arranging two circuits under test, the leaky-mode S-parameter can be de-embedded from the transmission line theory. Numerically, this method can avoid the problem of having to deal with a large-circuit-size structure when designing a leaky-wave antenna. The propagation constants of the leaky modes calculated by the full-wave spectral domain approach, and the experimental results of an aperture-coupled leaky-wave antenna are used to confirm the parameters extracted by this proposed technique. They all show good agreement. With appropriate modifications, the method can also be extended to extract and define the characteristic impedance of a leaky-wave antenna.</p>
<b>RESULT</b>	<p><b>Keywords:</b> S-Parameter [77]; Antenna [17]; Extraction technique [16]; Method [16]; Technique [9]; Characteristic impedance [8]; Experimental result [7]; Transmission line theory [7]; Aperture [5]; Aperture-Coupled[5]; Circuit [5]; Modification [4]; Paper [3]; Design [3]; Agreement [3]; Propagation [3]; Designing [2]; Arranging[2]; Deal [1]</p> <p><b>Descriptors:</b> S-Parameter; Antenna; Characteristic impedance; Experimental result; Transmission line theory; Agreement</p> <p><b>Non-Descriptors:</b> Extraction technique [16]; Method [16]; Technique [9]; Aperture [5]; Aperture-Coupled [5]; Circuit [5]; Modification [4]; Paper [3]; Design [3]; Propagation [3]; Designing [2]; Arranging [2]; Deal [1]</p> <p><b>Other possible terms:</b> leaky-mode S-parameter [51], proposed method [19], proposed technique [19], leaky-wave antenna [12], arranging circuits [12], experimental results of an aperture coupled leaky-wave [11], aperture-coupled leaky-wave [11], parameters extracted by this proposed technique [9], designing leaky-wave antenna [5], good agreement [4], efficient design [2], extraction technique for efficient design [2].</p> <p><b>Classification:</b> spectral range analysis, antenna design</p> <p><b>Unknown words:</b> numerically</p>

The human indexer can now change the list of descriptors if necessary by adding entries from the set *Other possible terms* (identify gaps in the thesaurus) or delete terms.

Within the list of *Unknown words* not only spelling errors but also new terms (including abbreviations) are listed for which the thesaurus developer has to decide whether they should become terms of the thesaurus or not.

### 3.4. Multilingual Indexing and Classification

Adding a new language requires an appropriate linguistic analysis for the new language(s), and either a corresponding bilingual thesaurus or if such a thesaurus is not available, a conversion dictionary which assigns terms of the new language to German or English terms. Mpro, the basic tool of the linguistic processing is currently available for 12 languages.

If none of these resources are available a bilingual thesaurus can be generated by using transfer lexicons for general language together with alignment tools applied to parallel or similar documents in the new language and in English or German. Due to the fact that English is the most widespread language, and therefore most of the bilingual resources will assign the native language into English, this language can act as an interlingua to generate other bilingual thesauri, for instance a German-Swedish thesaurus, taking a German-English and a Swedish-English dictionary

## 4. A Prototypical Application

The proposed approach has already been proven in a small bilateral project between IAI and FIZ Technik, Frankfurt, a bibliographic database producer, 500 German abstracts were automatically indexed using the FIZ thesaurus. The results were manually evaluated by comparing the manual indexed terms with the automatically computed indexing terms. The test results showed a good indexing quality, i.e. the ratio of automatically assigned terms to intellectually assigned terms was better than 70 %. Taking into account the higher speed, a document of an average size (1,5K) can be indexed and classified in approx. 25 seconds compared to 10 to 15 minutes a human indexer needs and the lower costs (using the automatic indexing tool the human indexer can process more documents in the same time) this small prototype has shown that such a tool can contribute to a better and more effective production cycle and therefore to a better market position for target user groups. It contributes also to a better consistency of the data and reduces the translation work for the indexing of foreign language documents.

## 5. References

1. **Bakel van, Bas.** *Modern Classical Document Indexing.* In: Proceedings of SIGIR'98.
2. **Bradshaw, Jeffrey M.** *Software Agents.* MIT Press, Cambridge, Mass., 1997.
3. **Dumois, Susan.** *Latent Semantic Indexing: TREC-3 Report.* Proceeding of the Third Text Retrieval Conference, 1994.
4. **Landauer, Thomas K., Michael L. Littman.** *A statistical method for language-independent representation of the topical content of text segments.* Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, 1990.
5. **Maas, Heinz Dieter.** *Multilinguale Textverarbeitung mit Mpro.* In: G. Lobin et al. (eds.): **Europäische Kommunikationskybernetik heute und morgen.** KoPäd, München, 1998.
6. **Maas, Heinz Dieter.** *Thesaurus als Wissensbasis für Begriffszerlegungen.* Information. Proceedings, Friedrich-Schiller-Universität Jena, 28. bis 30. September 1993.

7. **Plaut, Christian, Barbara A. Norgard.** *An Association Based Method for Automatic Indexing with a Controlled Vocabulary*, Technical Paper, University of California at Berkeley
8. **Salton, Gerald.** *Automatic Text Processing*. Addison-Wesley Publishing, 1989.
9. **Preissuchen** – WWW –Suchmaschinen, Kataloge und Metasucher im Vergleich, in: c't 23/99, pages 162ff.