Bilingual Indexing for Information Retrieval with AUTINDEX

Nübel, Rita[1]; Pease, Catherine[1]; Schmidt, Paul[2]; Maas, Dieter[1]

IAI[1]
Martin-Luther Str. 14, D-66111 Saarbrücken, Germany
{rita,cath,dieter}@iai.uni-sb,de}

FASK Universität Mainz[2]
An der Hochschule 2, Germersheim, Germany
schmidtp@usun2.fask.uni-mainz.de

Abstract

AUTINDEX is a bilingual automatic indexing system for the two languages German and English. It is being developed within the EU-funded BINDEX project. The aim of the system is to automatically index large quantities of abstracts of scientific and technical papers from several areas of engineering. Automatic indexing takes place using a controlled vocabulary provided in monolingual and bilingual thesauri. AUTINDEX produces for a given abstract a list of descriptors as well as a list of classification codes using these thesauri. It also allows for free indexing - indexing with an unrestricted vocabulary (delivering so called 'free descriptors´). These free descriptors are used to enhance and extend the thesauri. The bilingual AUTINDEX module indexes German abstracts in English and vice versa.

1. Introduction

The paper describes the AUTINDEX system which automatically indexes and classifies German and English texts. The AUTINDEX prototype has been further enhanced and adapted within the BINDEX[1] project with the goal of having a near-to-market system that can be integrated into existing production systems at the users' sites. The users in the BINDEX projects are FIZ Technik (Frankfurt a. M., Germany) and IEE/INSPEC (Stevenage, UK). FIZ Technik produces bibliographic documentary units (documents) which are indexed and classified on the basis of the FIZ Technik thesaurus and a corporate classification system. Documents to be indexed and classified are in English (> 70%) and German. Indexing is always done in German. INSPEC is recognised as the leading supplier of services in English, providing access to published literature in physics, electronics, and computing. The INSPEC database currently contains records for over six million scientific and technical papers and is being increased by around 330.000 papers per year. AUTINDEX[2] shall support human indexing of these growing amounts of documents.

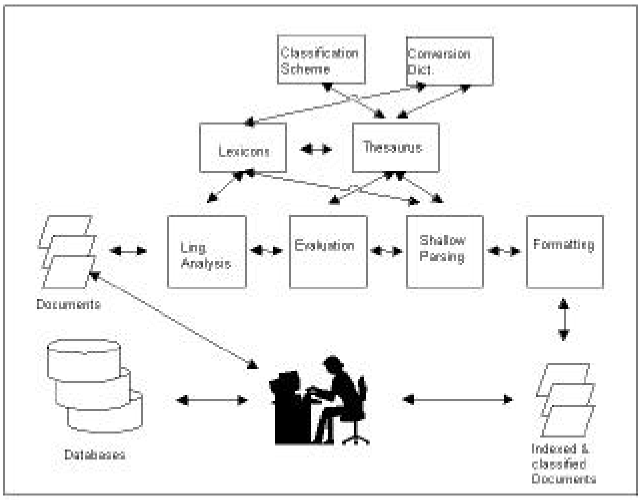
AUTINDEX is an NLP application which operates on the morpho-syntactic analysis of a document. It provides two types of indexing *at the same time*, namely free indexing which is purely based on the linguistic analysis, and controlled indexing which includes additional checking against a thesaurus for the calculation of key words. Additional resources like bilingual dictionaries and bilingual thesauri have been integrated for bilingual indexing, classification and free indexing. At the beginning of the project, the German module was the most advanced. The English and the bilingual English-German modules existed as mock-ups only and had to be extended considerably according to the user requirements. In this paper, stress is put on the development and integration of linguistic resources that are used in the indexing and classification procedure. Furthermore, the evaluation strategy for AUTINDEX and its results are described.

The remainder of this paper is organised as follows: Section 2 describes the NLP components of AUTINDEX. Section 3 describes the linguistic resources - general purpose and user-specific resources - which have been integrated in the system. In section 4 the evaluation strategy and the results are discussed, section 5 presents some related work, and in the last section we summarise our findings and discuss some future prospects.

2. The AUTINDEX 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.

The illustration below shows how the various components of AUTINDEX interact and which resources they use.



[1] BINDEX has been funded by the EU in the IST-Programme (IST- 1999-20028). The project started in November 2000 and was completed in April 2002.

[2] The AUTINDEX demonstrator is available under http://www.iai.uni-sb.de/bindex/home.html

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 here means the identification of keywords (thesaurus concepts) and free terms (candidates for thesaurus concepts) from the content of a document. 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.

AUTINDEX indexing and classification operates on the output of a morpho-syntactic analyser *MPRO* which has been developed at IAI, and on subsequent shallow parsing. *MPRO* performs the following actions:

- Word form identification
- Tagging
- Homograph resolution

These will be shortly described in the following subsections.

2.1. Wordform Identification & Tagging

Wordform identification recognises sentence boundaries, single words and fixed expressions including lexicalised multiword units such as *Schritt für Schritt* (step by step) or *Konstruieren-Herstellen-Erproben-Zyklen* (construction - production - testing cycles). A sequence of up to 10 units is looked up in the lexicon for fixed expressions. The output of this process is a feature bundle as shown below in (1)

(1) {string=W-Form, c=WD, sc=CAT, lu=CIT-Form, ...}.

It represents information about the wordstring (string=), the syntactic category (c=), the syntactic subcategory in case of e.g. function words, and about the normalised string (lu=). If the dictionary look-up produces more than one result, all of them are output. They are represented as a sequence of attribute-value pairs as exemplified in (2). Here, the analysis of *Die Frauen* (the women) produces two readings for the string *Die*, since it can be either a determiner or a relative pronoun (2):

(2) {ori=Die,pctr=no,lw=no,gra=cap,c=w,**sc=art**,**spec=def** , ...}
{ori=Die,pctr=no,lw=no,gra=cap,c=w,**sc=rel**}

Additional information includes e.g. the graphical representation, that is, if a word is written with a small letter or a capital letter at the beginning (gra=), if it is the last word of a sentence (lw=), or if it is followed by a punctuation sign (pctr=). As a matter of fact, wordform identification, tagging and lemmatisation take place at the same time, i.e. are collapsed into one analysis process with subsequent analysis steps resulting in one output file. They are described here separately for the sake of clarity.

Tagging and lemmatisation are based on a morpheme dictionary that contains (for German) around 52,000 entries, and for English around 45,000 entries. Together with a morphotactic module, strings are segmented into morphemes, then these morphemes are looked up in the morpheme dictionary which contains the information for morpheme combination and the morpheme's inherent properties. In order to avoid overgeneration and nonsense morpheme combinations, there is a stop list that contains prohibited words. E.g. the noun *Mitgliedschaft* can (theoretically) be segmented into *mitglied* (member) and *schaft. Schaft* is ambiguous because it is either a noun (shaft) or a derivation morpheme (-ship). However, the noun reading of s*chaft* is not appropriate here. In order to avoid this analysis, there is a feature introduced (rn=) that prohibits its analysis as the rightmost intraword. Hence, the analysis of s*chaft* as a noun is discarded (3):

(3) {string=mitglied,lu=mitglied,uni=yes,c=n,fuge=s|er|0,g=n,f=fs_es-fer,n={schaft=coll},s= agent&ano, **rn=schaft**}.

As illustrated in (3), the entries are also annotated with semantic information *s=agent&ano*. This semantic typing is exploited for indexing and free indexing as will be described later in this paper. After the segmentation process, the morphemes are concatenated. The output of word identification, tagging and lemmatisation then looks like the following:

(4) {wnra=1,wnrr=1,snr=1,ori=Mitgliedschaft,pctr=no,last=yes,pctl=no,offset=15,lw=yes,gra=cap,c=noun,endung=0,g=f,s=coll,t=mitgliedschaft,cs=n,ds=mitglied~schaft,ls=mitglied,ss=coll,lng=germ,w=1,ew=1,lu=mitgliedschaft,ts=mitgliedschaft,ehead={case=nom;gen;dat;acc,nb=sg,g=f,infl=weak;strong;null},saw=&n}

The *ori* attribute contains the string as it occurs in the text, *lu* is the citation form. The semantic analysis of *Mitgliedschaft* is output with the value *coll* (collective), and *schaft* occurs as derivation morpheme only (*ds=mitglied~schaft*). Gender and number are identified, and case has been left underspecified.

2.2. Homograph Resolution

Lemmatisation and tagging with *MPRO* takes place at word level only. Thus, ambiguities remain. A homograph reduction module partially resolves the remaining ambiguities. It consists of a set of rules that evaluate the word contexts on the basis of word order regularities and lexical information available from the dictionary. Sequences like *Er hat diskutiert* (he has discussed) produce multiple analyses for the past participle *diskutiert* (5c) since it can also be analysed as a present tense (2nd person plural (5b) or 3rd person singular (5a)) as in e.g. *Er diskutiert* (he discusses / is discussing):

(5)
a) {wnra=3,wnrr=3,snr=1,ori=diskutiert ,c=verb,vtyp=fiv,tns=pres,nb=sg,per=3, ,lu=diskutieren,...}
b) {wnra=3,wnrr=3,snr=1,ori=diskutiert, ,c=verb,vtyp=fiv,tns=pres,nb=plu,per=2, lu=diskutieren,...}
c) {wnra=3,wnrr=3,snr=1,ori=diskutiert, ,c=verb,vtyp=ptc2, lu=diskutieren,...}

The disambiguation module eliminates the analyses (a) and (b). However, ambiguities that cannot be resolved reliably remain and have to be coped with by subsequent modules.

2.3. Shallow Parsing

The shallow parsing component resolves remaining ambiguities. It also reliably identifies noun phrases (NPs). It determines the subject and the finite verb of a sentence and determines agreement. It consists of a number of phrase structure rules split into subgrammars which are successively applied (procedural processing). Subcategorisation information is not included in the dictionaries, thus no deep syntactic analysis is performed which means that not all ambiguities e.g. occurring with nominal elements (case ambiguity etc.) are resolved. This is, however, not relevant for subsequent indexing processes.

2.4. Indexing and Classification with AUTINDEX

AUTINDEX combines intellectual indexing with automatic free and controlled indexing, and it also classifies documents. Classification means determining the technical domain a document belongs to and is a kind of conceptual indexing.

2.4.1. Controlled Indexing with AUTINDEX

Controlled automatic indexing takes place in two major steps: first, an uncontrolled indexing takes place, using the above described NLP techniques to produce an unambiguous representation, and to identify multiword terms. The multiword terms and their syntactic variants (*Kostensenkung* versus *Senkung der Kosten* (cost reduction)) are identified using the shallow parser and additional sets of grammar rules which identify compound patterns and their variants. To calculate the keywords, AUTINDEX uses a statistical function based on frequency. It calculates a weight that is assigned to nouns depending on their semantic type. This means that frequency does not relate to simple word counts but it is based on the frequency of the semantic types of words (nouns including compound parts) occurring in a document. For English compounds which are mostly multiword units the semantic types of each word of the multiword unit are taken into account. The current inventory of semantic types assigned in the morpheme dictionaries amounts to 140. The result of the weighting is a set of key words which all belong to the semantic classes that have been calculated as most frequent classes.

In the next step these key words are checked against a thesaurus provided by the users FIZ Technik (German) or INSPEC (English). Here, also the classification codes are calculated. No additional knowledge base is needed. Also, the hierarchical structure of the thesaurus, i.e. hyperonym relations as well as synonyms are used in order to calculate the set of *descriptors*. Descriptors are valid thesaurus denotations, and most of them are annotated with their corresponding classification code. The FIZ Technik classification scheme consists of 348 classes which are also organised hierarchically. The INSPEC classification scheme is much more fine-grained and consists of 3292 classes. In case ambiguous descriptors (socalled bracketed descriptors, e.g. *Zahn (dens;Maschinenteil)* (tooth (dent;machine part)) are calculated, the classification can be used for disambiguation purposes. In this case, the descriptors are output if they belong to the calculated classifications, e.g. *3BZB* (biological basics in medicine) or *3MB* (machine elements). The thesauri have been formatted into a format that can be processed by the linguistic analysis components of AUTINDEX.

2.4.2. Free indexing with AUTINDEX

Free indexing means that socalled free descriptors having the following structures are calculated under the conditions that they have not yet been calculated as thesaurus terms:

- Compound nouns
- NPs of the type Adjective-Noun
- Simple nouns annotated with selected semantic types

The calculation of free terms is based on the linguistic resources as described here. No other resources are necessary, neither for monolingual nor for bilingual indexing.

The final output of AUTINDEX is a file with structured indexing information organised in the fields as defined by the project partners. Below, we illustrate such an output file.

Processed File:
06710936.sgm
Title:
Enhancing far infrared image sequences with model-based adaptive filtering
S-Title:
K-Title:
Z-Title:
Chinese Journal of Computers
Z-Code:
Abstract:
Two enhancement algorithms, spatial and spatio-temporal homomorphic filtering (SHF and STHF) were proposed by Highnam et al. (1997) to enhance far infrared images based upon a far infrared imaging model. This paper proves that the enhanced results with SHF are in general smoother than those with STHF, although STHF may reduce the processing time greatly in comparison to SHF. Based on this conclusion, an adaptive spatio-temporal homomorphic filtering algorithm, ASTHF, is proposed. The adaptive factor of ASTHF is also discussed in detail to obtain the tradeoff between the smoothness and the convergence.
Section 1:
Reference Titles:
Manually assigned descriptors:
adaptive filters; filtering theory; image enhancement; image sequences; infrared imaging
Automatically assigned descriptors:
adaptive filters[100]; infrared imaging[55]; image sequences[53]; modelling[22]; image enhancement[4]
Matching descriptors:
image enhancement; image sequences; infrared imaging; adaptive filters
Recall:
80.00%
Precision:
80.00%
Consistency:
66.67%
Manually assigned free terms:
far infrared image sequences; model-based adaptive filtering; adaptive spatio-temporal homomorphic filtering algorithm; smoothness

Automatically assigned free terms:
Highnam et; Highnam et al; adaptive factor; enhancement algorithm; et al; far image; far image sequence; far infrared image; far infrared image sequence; far infrared imaging; filtering algorithm; homomorphic algorithm; homomorphic filtering; homomorphic filtering algorithm; infrared image; infrared image sequence; model-based adaptive filters; processing time; spatio-temporal filtering; spatio-temporal homomorphic filtering
Manually assigned classification:
B6135; B6140B; C5260B; C1260S
Automatically assigned classification:
A4230V[100] (image sequences;image enhancement); B6135[100] (image sequences;image enhancement); C1250M[100] (image sequences;image enhancement); C5260B[100] (image sequences;image enhancement); C1260S[93] (image enhancement;adaptive filters); C5260D[92] (image sequences); A0720[86] (infrared imaging); A0762[86] (infrared imaging); A4280Q[86] (infrared imaging); B1270[86] (adaptive filters); B6140B[86] (adaptive filters); B7230G[86] (infrared imaging); B7730[86] (infrared imaging); C5240[86] (adaptive filters); C1220[19] (modelling)
Matching classification:
B6135; C5260B; B6140B; C1260S
Recall:
100.00%
Precision:
26.67%
Consistency:
26.67%
Countries:
Unknown words:
ASTHF; Highnam; SHF; STHF; al; et

Figure 2: Exemplary AUTINDEX output file.

For evaluation purposes, the manually assigned descriptors, the classification and the free descriptors or free terms are copied into the index file. The values for the quality parameters recall, precision and consistency are based on the comparison between manually and automatically assigned descriptors and classification codes. The numbers in square brackets in the classification as well as the descriptors fields indicate the weighting. The higher the number that has been calculated, the higher the appropriateness of a descriptor or a classification.

Bilingual indexing in AUTINDEX refers to two variants:

- indexing German documents with INSPEC specific English descriptors and classifications
- indexing English documents with FIZ Technik specific German descriptors and classifications.

The processing steps are those described above for monolingual indexing plus subsequent translation processes. Additional adaptations of the resources were necessary in order to perform the bilingual indexing for both user-specific configurations. The strategy of bilingual indexing consists in indexing a document in the document language first, and then to translate the calculated information on the basis of additional resources. The adaptation, enhancement, and integration of the linguistic resources of AUTINDEX for both monolingual and bilingual indexing are described in the next section.

3. Linguistic Resources in AUTINDEX

The major resources for both free and controlled indexing are the morpheme dictionaries. Additionally, the user specific thesauri are converted and used for controlled indexing. For the calculation of free descriptors no other resources are needed: they purely stem from the linguistic analysis of the AUTINDEX system. Additional resources like synonym lists and lists of semantic decompositions have been integrated in the system. They will be described here together with other resource extensions. The main resources for the bilingual module were in a first experimental phase the IAI transfer dictionaries English-German. As additional requirements came up, the translations of descriptors had to be "controlled" through the thesauri, i.e. they had to be translated into corresponding English or German descriptors according to FIZ Technik or INSPEC. Additionally, a conversion dictionary that maps English terms onto German FIZ Technik descriptors and free terms was adapted and integrated in the AUTINDEX system.

3.1. Resources for German Indexing

At the beginning of the BINDEX project, the German AUTINDEX module was the most advanced. It builds on the following linguistic resources:

- German morphological dictionaries
- a bilingual thesaurus from FIZ Technik consisting of 49,703 entries
- a list of forbidden words consisting of 86 entries which bloc the respective descriptor, since it is considered as being too general e.g. *Ausführung* (model;execution) or *Ereignis* (event)
- a word form dictionary for lexicalised compound in order to prevent inappropriate decompositions for descriptor calculation, e.g. *Kunstharz* (synthetic resin) should not be decomposed into *Kunst* (art) and *Harz* (resin) which might lead to unwanted descriptors.

In the course of the evaluation cycles, additional requirements were specified by the users and the resources were adapted and extended accordingly.

3.1.1. Extending the morpheme dictionary

The existing morpheme dictionary for German covers nearly 99% of the German language. It consists of around 52,000 entries and is a constantly growing resource. For AUTINDEX, we had to add e.g. chemical formulas like *(Bi&cmPb)2Sr2Ca2Cu3O(x)* or other chemical names like *soda Lye.* Furthermore, the existing list of some hundred geographical names like cities or regions has been extended and the final version consists of 7600 entries. These are used for the calculation of country codes (see the field "Countries"), e.g. *Afrika C60AFR*.

3.1.2. Adaptation of the German FIZ Technik thesaurus

The original German thesaurus consists of 49,703 descriptors. It can be seen as a defined documentation language. However, this documentation language contains structures that are not just a partial set of the German language, as shown in the following example entries:

(6) Stahl (nach Anwendung) (steel related to applications)

(7) Steg (Planetengetriebe) (planetary gears)

Constructions like (6) or (7) are no natural German expressions. Thus, indexing can be seen as a sort of translation where German expressions are mapped onto expressions of the thesaurus language. Furthermore, the FIZ Technik thesaurus terms are written according to the old German orthography and have to be rendered as such. The morphosyntactic analysis therefore has to convert the new orthographic structures into the old version. The strategy was thus to produce besides the original thesaurus another "German-German transfer thesaurus" on the basis of the original thesaurus. This transfer thesaurus maps German expressions onto German thesaurus expressions in order to be able to map e.g. *Steg* onto the construction in (7), since *Steg* alone is not a descriptor.

3.1.3. Development of Additional Resources

Global synonyms. A list of global synonyms containing 252 entries was provided by FIZ Technik on the basis of evaluation results. These synonyms are matched against any part of any thesaurus entry including compounds. Below we illustrate a global synonym class consisting of two synonyms:

(8) {t=abfall;müll,gs=m;m}.

If in the text the noun *Müllwirtschaft* (waste management) occurs -- which is not a descriptor, the compound part *Müll* is exchanged with its synonym *Abfall*, which results in *Abfallwirtschaft* -- which is a descriptor.

Semantic decompositions. Based on test results FIZ Technik provided a list of 422 socalled semantic decompositions and derivations (string=A_(D)) which are mapped onto thesaurus terms (besser=B). Thus, non-descriptor expressions (NPs) are mapped onto one or more descriptors. The full-fledged list consists of 1547 mappings. In (9) and (10) we illustrate two example mappings:

(9) {string=abbaubares_Tensid,c=rename,besser=
Abbaubarkeit)

(10) {string=Pn-struktur,c=rename,besser=
pn-Übergang}

Semantic decompositions and global synonyms have been integrated in order to increase the recall and precision, and to reduce "noise", i.e. irrelevant descriptors on the other hand.

Similarly, a list of so called non-term parts was integrated in the linguistic analysis. It consists of 343 entries which have been defined as invalid parts of possible term candidates (multiword units). For example, the participial adjective *getestet* (tested) belongs to this list, and in an indexed text it may be part of a multiword unit, consisting of two adjectives and a noun (underlined here):

(11) Hier werden getestete rote Blutkörperchen analysiert.
(Tested red blood cells are analysed here.)

In the German module, the calculation of multiword units as free terms is output in the corresponding field. As defined, these free terms consist of multiword units which may have the pattern *Adjective-Noun*. The linguistic analysis identifies *getestete rote Blutkörperchen* as candidate construction of a free term. Then this construction is checked against the list of non-term parts in order to find out whether the first word in this construction is on the list. If so, this word is eliminated (iterative operation). Constructions which consist of two words or more after this operation are normalised and then output. In the above described example, the calculated free multiword term will be *rotes Blutkörperchen*, but not *getestetes Blutkörperchen*.

3.2. Resources for English Indexing

For English monolingual indexing we started with a morpheme dictionary of approximately 37,000 morphemes. The English INSPEC thesaurus to be integrated for controlled indexing has the size of 15,000 entries 8,000 of which are descriptors, most others are synonyms. Additional user-specific resources were developed and integrated.

3.2.1. Global Synonyms

A list of synonyms containing 35 entries was prepared on the basis of test results which can be matched against any part of any thesaurus entry, see some exemplary entries below (12):

(12) analysis → estimate
analysis → estimation
encryption → encryptography
encryptography → encryption

3.2.2. Derivational Synonyms

The thesaurus resources were also enhanced by including appropriate derivations of thesaurus entries for both descriptors and descriptor synonyms. This means that approved derivations are also cross-referenced with thesaurus entries. Approved derivations number approximately 3,000 and are based on the following types:

- agentive derivation → nominalisation
  air pollutant → air pollution
- relational adjective → noun
  photochemical → photochemistry
  biomagnetic → biomagnetism
- noun → relational adjective
  algebra codes → algebraic codes
- past participle → nominalisation
  polluted air → air pollution
  transmitted DC power → DC power transmission
- verb → nominalisation
  (to) calculate augmented plane waves → augmented plane wave calculation
- verb → agentive derivation
  (to) control cerebellar model articulation → cerebellar model articulation controller

3.3. Resources for Bilingual Indexing

The task of the bilingual module is the following

- Automatic text language recognition
- Indexing of German documents in English and vice versa
- Translation of indexing information into either English or German depending on the input text language

Additionally, English indexing of German texts had to be based on INSPEC resources (INSPEC configuration), whereas German indexing of English texts had to be based on FIZ Technik resources (FIZ Technik configuration).

This made a number of adaptations and extensions of linguistic resources necessary.

3.3.1. English to German Indexing

The task of the AUTINDEX system for the FIZ Technik configuration consists in indexing an English document with German descriptors and classifications according to the FIZ Technik standard and coding norms. The processing steps for bilingual English to German indexing are identical with those of the monolingual English indexing module except for two important features:

- Additional production and use of English FIZ Technik resources for the English indexing (first step)
- Translation of English indexing into German, using FIZ Technik resources (second step).

Additional production and use of English FIZ Technik resources. First, the FIZ Technik thesaurus which provides English translations of descriptors for 39.171 entries had to be "reversed", i.e. the English translations had to be extracted from the original thesaurus. Sometimes, two translations are associated with a German descriptor. In these cases, only the first translation was chosen in order to build the English FIZ Technik thesaurus. This English thesaurus was used for the automatic English indexing of English abstracts. For this, the English monolingual indexing module was used without further changes except the linking with the appropriate FIZ Technik resources.

The translation of the English descriptors was performed on the basis of the FIZ Technik thesaurus which was now available in English and in German. We automatically produced a transfer dictionary from these two thesauri. For the translation of the automatically assigned free terms we used the FIZ Technik conversion dictionary which maps English INSPEC terms onto German FIZ Technik descriptors (which are almost all included in the FIZ Technik thesaurus) or onto free terms. The number of entries in the conversion dictionary which map English terms onto German free terms is 106.210. Additionally, the IAI English-German transfer dictionary containing around 488.000 entries was used for the translation of free terms for which no translation was provided in the conversion dictionary.

Translation of English indexing information into German. For the translation of descriptors and free terms we integrated an already existing transfer component which has been developed at IAI. It is not a full-fledged transfer system with e.g. an analysis, transfer, and generation / synthesis component. However, for the purpose of the translation of descriptors and free terms the latter of which have relatively simple patterns (simple or compound noun; adjective + noun; noun + preposition + noun), no complex transfer operations were necessary.

3.3.2. German to English Indexing

The task of the AUTINDEX system for the INSPEC configuration consists in indexing a German document with English descriptors and classifications according to the INSPEC standard and coding norms. The processing steps for bilingual German to English indexing are identical with those for the monolingual German indexing module except for the following additional features:

- Additional production and use of German INSPEC resources for the German indexing (first step).
- Translation of manually assigned German FIZ Technik descriptors into English using INSPEC resources in order to enable comparison of manually assigned descriptors with automatically assigned descriptors.
- Translation of German indexing into English, using INSPEC resources (second step).

Additional production and use of German INSPEC resources. In contrast with the FIZ Technik thesaurus, which provides English translations for almost all of its German descriptors, the INSPEC thesaurus is monolingual. We used these data in order to produce adequate user-specific bilingual resources for the INSPEC indexing configuration.

First, we had to produce a German INSPEC thesaurus which has to provide descriptor entries which enable the calculation of INSPEC classifications. Additionally, a transfer dictionary was produced that translates the German INSPEC descriptors back into English INSPEC descriptors. In a first experiment the INSPEC thesaurus was translated with the IAI English-German transfer dictionary and the conversion dictionary. This seemed appropriate at first sight, since it is not important whether these intermediate German translations are true descriptors or not. The INSPEC English original descriptor was kept anyhow together with its German translation, in order to use it for the backtranslation into English. However, this strategy raised problems because it produced ambiguous translations which could not be disambiguated in the backtranslation process, so that the translation back into English INSPEC descriptors would produce wrong results. We therefore revised the strategy towards a more controlled backtranslation. The English INSPEC thesaurus was translated into German using the existing FIZ Technik bilingual resources, i.e. the English-German FIZ Technik "transfer thesaurus" (which had been produced already for the bilingual indexing in the FIZ Technik configuration as described in the former sections, see the example entries (10) and (11) above). Additionally, we also used the FIZ Technik conversion dictionary. The resulting German INSPEC thesaurus consists of 9762 entries.

The backtranslation of the German thesaurus into INSPEC English descriptors was performed on the basis of the two INSPEC thesauri. We automatically produced atransfer dictionary from these two thesauri, consisting of 8,092 entries *plus* 1858 entries which could be extracted from a bilingual word list with correspondences of FIZ Technik descriptors and INSPEC descriptors which had additionally been provided by FIZ Technik. This transfer thesaurus dictionary is relatively small (9950 entries), but it guarantees that there are no ambiguities left.

For the translation of the automatically assigned free terms we used the reversed (!) FIZ Technik conversion dictionary mapping German descriptors and free terms onto English terms. The number of entries in the reversed conversion dictionary is 89,018 entries. Additionally, the IAI German-English transfer dictionary containing around 488,000 entries was used for the translation of free terms for which no translation was provided in the reversed conversion dictionary

Translation of German indexing information into English. For the translation of descriptors and free terms we used the already described transfer component which has been developed at IAI and which uses the above described resources. In addition to the translation of the automatically assigned German descriptors into English, we also translated the *manually* assigned German descriptors into English (using the same transfer dictionary produced from the two INSPEC thesauri as described), in order to enable the automatic calculation of precision, recall and consistency, and to allow for a more straightforward human evaluation to be carried out at the user's site. We did not do this with the manually assigned free terms since it was not specified as user requirement. Additionally, it was of course not possible to compare the manually assigned German FIZ Technik classification with the automatically calculated INSPEC classification, since they are not 1:1 correspondences.

4. AUTINDEX evaluation

The basis of the AUTINDEX evaluation rounds were 552 German abstracts, and 807 English abstracts. At the end of each development phase, there was a testing round at the developer's site. When results were considered acceptable and in accordance with the goal of the development phases, the indexed abstracts were submitted to the users for official evaluation.

4.1. The evaluation strategy

The evaluation strategy consisted in assessing the quality of automatic indexing and classification in terms of precision, recall, and consistency, which were calculated automatically in the final version of the AUTINDEX system. The benchmark for these assessments was the human indexing and classification which was available for both German and English monolingual indexing as well as for the German indexing of English abstracts in the FIZ Technik configuration. For the INSPEC configuration of bilingual indexing however, this benchmark did not exist, since there were no German documents available that had been manually indexed with INSPEC English resources. Hence, this characteristic had to be simulated somehow: in order to allow for a comparison of descriptor calculation, the German FIZ Technik descriptors were translated automatically using the INSPEC German-English thesaurus which had been produced as described in the previous sections.

4.2. Evaluation Results

German monolingual indexing:

The results for monolingual indexing of German documents are the following:

Calculation of descriptors:

Recall:	*34%*
Precision:	*20%*
Consistency:	*15%*

Calculation of classification:

Recall:	*61%*
Precisiom:	*15%*
Consistency:	*14%*

English monolingual indexing:

The results for monolingual indexing of English documents are the following:

Calculation of descriptors*:*

Recall:	*37%*
Precision:	*30%*
consistency:	*20%*

Calculation of classification:

Recall:	*37%*
Precision:	*12%*
Consistency:	*10%*

Bilingual indexing German-English:

The results for indexing German documents with English INSPEC data are the following:

Calculation of descriptors:

Recall:	*34%*
Precision:	*13%*
Consistency:	*11%*

No values were calculated for the classification since the automatic assignment refers to INSPEC classification classes whereas the German manually assigned classification codes refer to FIZ data.

Bilingual Indexing English-German:

The results for indexing English documents with German FIZ data are the following:

Calculation of descriptors:

Recall:	*:26%*
Precision:	*13%*
Consistency:	*10%*

Calculation of classification:

Recall:	*50%*
Precision:	*17%*
Consistency:	*15%*

4.3. Evaluating the Evaluation Strategy

Taking the manual indexing was considered the most pragmatic and straightforward strategy given human resources and time constraints within the project, but it bears the risk of "losing" useful automatic indexes: it is not possible to automatically assess them, if they are *not* in the set of matching descriptors. Precision and recall was calculated only on the basis of these matching (intellectually and automatically assigned) descriptors and classification codes. Additionally, intellectual indexing was done on the basis of full texts, whereas for automatic indexing only the shorter abstracts were used, which obviously results in a lack of information. In a small-scale test of 30 indexed German abstracts, where human indexing was also based on short text versions only, recall increased to 46%. However, large retrieval tests would have been more useful here, but could not be carried out due to time restrictions.

5. Related Work

Commercial indexing systems such as CINDEX 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 (Dumois 1994). 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 (Plaunt, Norgard 1997) 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 (van Bakel 1998). 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 then 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.

6. Summary and Future Prospects

The aim of the BINDEX project was to further develop the AUTINDEX indexing and classification system on the basis of the users' requirements. A lot of enhancements and new developments were necessary for the English monolingual as well as the bilingual modules, since these existed as small-coverage mock-ups only when the project started.

The AUTINDEX system provides three stable modules which produce automatic indexing and classification using the a controlled documentation language. This feature is therefore the highly user-specific characteristic of AUTINDEX.

The AUTINDEX system *also* produces so called free indexing information (free terms), which is done completely independent of the controlled (thesaurus-specific) indexing and classification. Due to this characteristics, AUTINDEX can also be used for the free monolingual and bilingual indexing of *any type of content* available in English or German. This option will be interesting especially for users who do not yet use a thesaurus or something similar, because it supports them in building up such a resource. Additionally, extending the free indexing module to other language can be done quite straightforwardly, since no user-specific resources - including bilingual resources, have to be adapted and integrated. Since IAI's MPRO system and the corresponding lexicon resources including transfer dictionaries (mostly from to English or German) covers a number of European languages including French, Italian, Spanish, Portuguese, Dutch and Swedish, but also to a small extent Bulgarian, Russian and Greek, extending the free indexing component to other languages seems an attractive option.

Extending AUTINDEX to other users by integrating user-specific resources has also already been considered. Here, it will be important that (ideally) from the beginning the user's resources like the thesaurus provide an internal structure that can fully be exploited by the automatic indexing modules, e.g. ambiguities should be avoided.

7. Acknowledgements

We would like to thank Christoph Horschmann for valuable suggestions and comments.

BINDEX has been funded by the EU in the IST-Programme (IST- 1999-20028). The project started in November 2000 and was completed in April 2002.

8. References

[van Bakel 1998] van Bakel, Bas: *Modern Classical Document Indexing.* In: Proceedings of SIGIR 1998.

[Dumois 1994] Dumois, Susan: *Latent Semantic Indexing: TREC-3 Report*. In: Proceedings of the Third Text Retrieval Conference, 1994.

[Maas 1998] Maas, Heinz Dieter: Multilinguale Textverarbeitung mit MPRO. In: Lobin, G. et al.(eds): *Europäische Kommunikationskybernetik heute und morgen*. KoPäd, München 1998.

[Nübel & Pease 2002] Nübel, Rita and Catherine Pease: *Report on the Assessment and Adaptation of the BINDEX Resources*. BINDEX deliverable D3.1, Saarbrücken, March 2002.

[Plaunt, Norgard 1997] Plaunt, Christian and Barbara A. Norgard: *An Association Based Method for Automatic Indexing with a Controlled Vocabulary,* Technical Paper, University of California at Berkeley , 1997.

[Ripplinger, Schmidt 2000] Ripplinger, Bärbel. and Paul Schmidt: *Automatic Multilingual Indexing and Natural Language Processing*. In: Proceedings of SIGIR 2000, New Orleans.