SUPPORTING MULTILINGUALITY IN LIBRARY AUTOMATION SYSTEMS USING AI TOOLS

STEPHANOS MICHOS AND EFSTATHIOS STAMATATOS
Knowledge S.A., Patras, Greece

NIKOS FAKOTAKIS
University of Patras, Wire Communications Laboratory, Patras, Greece

Language barriers present a major problem in the effectiveness of resource sharing and in common access to the resources of libraries. In this paper we present the TRANSLIB system, which consists of an integration of both new and existing multilingual information tools. This system takes full advantage of some AI–based methods in order to provide multilingual access to library catalogues. Its main features include functionalities for searching in multiple languages, multilingual presentation of the query results, and localization of the user interface. TRANSLIB has currently been tested in existing medium–sized bibliographic databases. Evaluation results show a remarkable improvement in the search process and report high user friendliness and easy and low–cost maintenance and upgrade of the system.

Today's libraries have automated their everyday transactions such as acquisition, cataloguing, and circulation. An automated library appears to users as an online public access catalogue (OPAC) through which they can quickly search and obtain the desired information (Leeves, 1994). Such systems support bibliographic search according to author name, title, publication year, etc., and provide free text and keyword facilities. Hence, computer systems have replaced the librarian's role as intermediary between user and library, and expert systems that simulate this role providing general information to the user have been developed (Morris, 1992).

Moreover, the application of AI to information retrieval has recently attracted the attention of information scientists. Advanced applications of natural language processing (NLP), such as machine translation, have

Final version received November 1998.

This work was supported by a European Union grant under the contract LIB93–3038. The companies/institutions that contributed to the design, development, and testing of the presented system are: Knowledge S.A., University Carlos III of Madrid, Library of Spanish Agency of International Cooperation, Central Library of University of Patras, and Municipal Library of Patras. The views, opinions, and/or findings contained in this paper are those of the authors and should not be construed as an official EU (or other partners) position, policy, or decision.

Address correspondence to Stephanos Michos, Knowledge S. A., N.E.O. Patron–Athinon 37, 264–41 Patras, Greece. E–mail: smichos@knowledge.gr

Applied Artificial Intelligence, 13:679–703, 1999

0883–9514/99 $12.00 + .00

already improved the capabilities of several systems in this area (Gibb & Smart, 1991; Gibb, 1993). Translation–based methods for translingual information retrieval require the query be translated into the target language. In this case, full–fledged machine translation is not applicable. Experience shows that of three well–known translation approaches, which is dictionary–based term translation (DICT), example–based term translation (EBT), and example–based term–to–sentence (EBS), DICT presents the best solution (i.e., the one enhancing, recall but at a cost in precision) for queries consisting of isolated words, or at best, short phrases (Yang et al., 1997).

Other approaches in information retrieval are concerned with systems with cross–language functionality. A few of them, especially in trade transactions and Web–based information services, make use of a controlled language in order to improve the recall and precision rates (Lehtola & Honkela, 1996; Schütz, 1996). The majority of these systems utilize free text retrieval (Oard, 1997). The corpus–based approach, MIRTH (Zhang et al. 1997), supports within–language retrieval in English and Chinese and in specific areas, such as computing, linguistics and literature. The limitations of this system lie in the lack of a complete linguistic tool to guide translation, as well as in the absence of any maintenance system to deal with tasks such as add, insert, delete, update, replace, and sort links with their keywords. Knowledge–based information retrieval is used in Gilarranz et al. (1997), where the features of the EuroWordNet multilingual lexical knowledge base are described. However, although both approaches tackle the cross–language functionality of the retrieval, they do not consider as yet the localizability of their system by adopting, for instance, an internationalized methodology in their design.

In addition, libraries have not paid special attention to multilingual features of OPACs, despite the fact that it is a serious problem (Cousins & Hartley 1994). The National Library of Canada was one of the first libraries to offer multilingual access to its users, in the form of controlled bilingual, that is English and French, authority files (Buchinski et al., 1976). It is also worthwhile mentioning the ETHICS project that produced an OPAC with multilingual user interface and help screens and a subject index in three languages (French, German, and English) (Hug & Noethinger, 1988).

So far, multilinguality in OPACs has been dictated by the needs of a specific multilingual community and restricted to the provision of bilingual or multilingual lists of controlled terms such as controlled authority files, subject headings, and thesauri (McAllistair, 1987; Slater, 1991; Butcher, 1993). Multilinguality as a potential problem for the common user of the library was only explicitly investigated in the NORDINFO survey (Pasanen–Tuomainen, 1992a; Pasanen–Tuomainen, 1992b). According to the results of this survey, multilingual access to online catalogues can improve remarkably the quality of the provided services.

Similar user surveys were undertaken by the Central Library of the University of Patras, the Library of the Spanish Agency of International Cooperation, and the Municipal Library of Patras (Synellis, 1995). The main outcomes of the above surveys have been as follows:

- Searching in a conventional OPAC by using controlled or uncontrolled terms is time consuming, requiring repeated effort from the user who in a lot of cases remains unsatisfied.
- The libraries express their interest in a multilingual access tool, because they understand that multilinguality poses problems to their users as well as to the (widely desired) interconnection with libraries from other countries.
- The translation tool, if any, should be able to retrieve information from the title, subject, and author fields.
- An important indication is that the specific national conditions should be taken very seriously into account, especially with respect to dialects and unofficial languages. A very good example is the relatively high interest of the Spanish library in the Catalan language.

These surveys were aimed at an analysis of the attitude of the users toward the OPAC they used and investigation about their eventual need for a multilingual tool. The results of these surveys are essentially independent of the sex, educational status, familiarity of computer use, and frequency of the use of the OPAC by sample users. The analysis of the results was very illuminating:

- approximately 75% of the users were either moderately satisfied or dissatisfied with the results of their searches in the OPAC;
- over 80% of the dissatisfied users noted that the hits were not in the user's native language or that the search was based only on a single language;
- approximately 70% of the users were interested in bibliographies written in foreign languages;
- 65% of those who used books in only one language considered being able to search in multiple languages either useful or very useful.

On the other hand, despite the adoption of AI by recent significant library automation systems, NLP techniques have not yet been applied to the development of real multilingual interfaces and tools for the translation of keywords, titles, or abstracts (Fluhr et al., 1996; Kikui et al., 1996). Furthermore, previous surveys have also shown that 64% of the users consider a potential translated title presentation in their native language very useful, whether or not they use one or more languages (Synellis, 1995).

In this paper we present TRANSLIB, a multilingual interface system to library automation systems. It takes full advantage of tools such as bilingual dictionaries, conversion tables, terminology lexica, intelligent thesauri, and simplified translation tools in order to support multilingual access to library catalogues. This system stemmed from the integration of new and existing multilingual information tools and has been tested in existing medium–sized bibliographic databases in Greece and Spain. The tools are characterized by a high degree of modularity and user–friendliness that allow easy and low–cost maintenance.

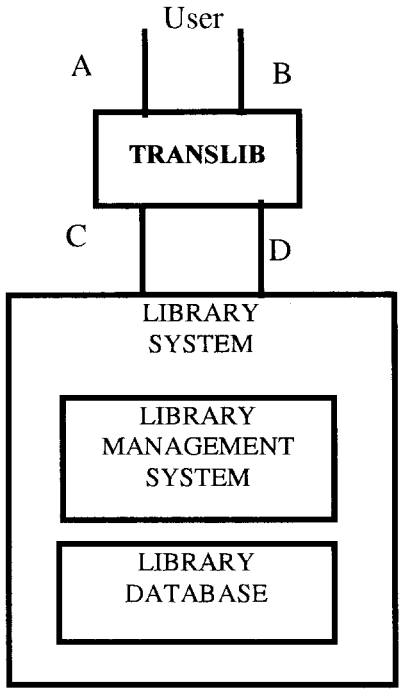
The following section contains an overview of the TRANSLIB system, describes briefly its architecture and gives its basic features. Section 3 describes in detail the multilingual resources of the presented system and section 4 includes some evaluation results that illustrate its impact on the improvement of library services. Finally, in section 5 some conclusions are drawn and the steps to be taken to make TRANSLIB a marketable product are discussed.

OVERVIEW OF TRANSLIB

The global TRANSLIB perspective is shown in Figure 1. The user interacts with the TRANSLIB module by choosing the languages to be used (i.e., input, message, search, and output languages) as well as the search fields, the search key topics, and the type of searching to be initiated (arrow A). TRANSLIB combines all this information to support multilingual access to the library system, by sending it a query for each key topic (arrow C) and receiving from it the records that match the users' query (arrow D). Finally, TRANSLIB presents the query results to the user in the preferred language as well as the predefined search field labels, information windows, and all messages (arrow B). If the query results are empty, TRANSLIB allows the user to repeat the search using synonyms, supercategories, or subcategories of the search keywords.

TRANSLIB is an OPAC that provides the users with a capability for multilingual access to library catalogues. TRANSLIB is fully implemented, runs under Windows 95, and currently supports three languages, that is English, Greek, and Spanish (Stamatatos et al., 1997). These three languages are considered to be sufficient to both demonstrate the feasibility of this type of multilingual access and to highlight potential problems in the future adoption of additional European or non–European languages. An outline of the system is given in Figure 2.

TRANSLIB offers bibliographic information retrieval, whether from a local library database or from remote databases, through servers that support the Z39.50 standard, that is, a standard specifying a client/server–based protocol for information retrieval that was originally proposed for use



A: Input, message, search, output language(s)

B: Presented query results, fields, labels, messages

C: Search topic query

D: Received records

FIGURE 1. TRANSLIB perspective.

with bibliographic information (ANSI/NISO Z39.50, 1995). One of the main characteristics of TRANSLIB is its ability to be easily integrated into the majority of library automation tools. In order to achieve this objective, we have had to consider seriously the following problems:

- all library automation systems do not use the same type of database (e.g., relational, nonrelational, etc.);
- there will be libraries that will not give access to their databases to developers' groups so that they can be connected with the TRANSLIB system (license problems).

Our solution has been to integrate a Z39.50 client tool into TRANSLIB (see Figure 3). Thus, the TRANSLIB interface can now connect (via its embedded Z39.50 client module) with any library automation system that

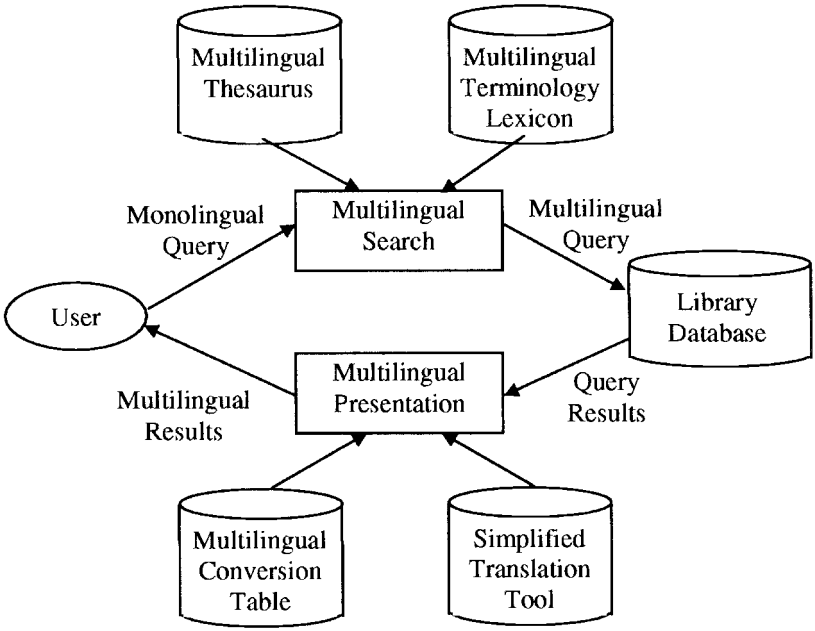


FIGURE 2. An outline of the TRANSLIB system.

supports the Z39.50 protocol. All other processes, such as retrieving bibliographic information in any language and translating the titles and keywords into English are also supported by the new TRANSLIB with the Z39.50 client module.

TRANSLIB comprises two basic tools:

i. *the multilingual search tool* that supports the retrieval of multilingual bibliographic information, and
ii. the *multilingual presentation tool* that allows multilingual presentation of the retrieved information.

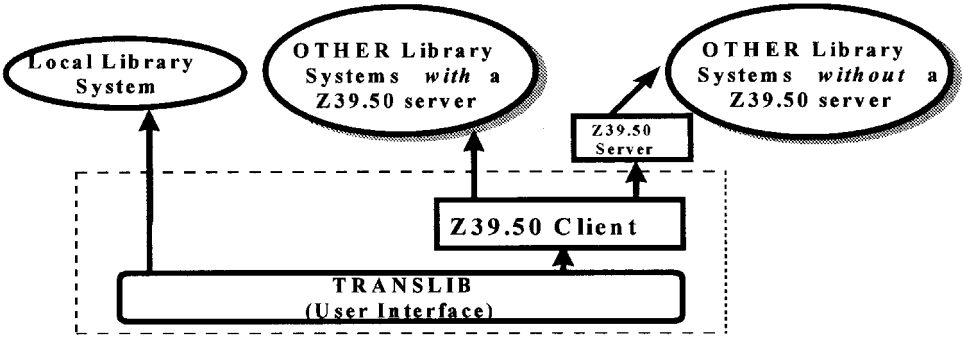


FIGURE 3. The TRANSLIB system with the Z39.50 client tool.

The following subsections illustrate the features of these tools. By taking advantage of them, TRANSLIB offers the following facilities in contrast to up–to–date library systems (Leeves, 1994):

- Localization of user–interface (i.e., English, Greek, or Spanish) is supported.
- The user is able to enter a query in his/her native language and search for entries that match this query in all the possible languages for which there is at least one entry in the current database.
- The retrieval of all the entries for a given query as well as for its synonyms in any language is possible. Furthermore, the user is able to indicate the type of search, which can be based on narrower and/or broader terms of the query.
- Finally, translation of the retrieved books/journals titles and keywords in any language is provided in order to give the user an idea of the content of these books/journals.

Multilingual Search Tool

This tool enables the user to enter the query in the language (s)he prefers and select the languages in which the matching entries have to be found. Essentially, this tool performs a conversion from monolingual to multilingual queries. The user can select the input language and the search languages (s)he prefers:

- *Input language*: the language used to introduce the search criteria, that is, the title, author, publication year, etc.
- *Search language*(*s*): one or more languages in which the user wishes to find some documents matching his queries.

Furthermore, the user is able to determine the *search depth* by selecting to search with synonyms, and/or narrower terms, and/or broader terms of the search criteria. The multilingual search tool utilizes:

- a *multilingual terminology lexicon*, allowing the search of keywords in several languages, and
- a *multilingual thesaurus*, enabling sophisticated bibliographic information retrieval by means of synonyms, narrower and broader terms.

An example of broader and narrower terms of some keywords is shown in Table 1. The search criteria can be specified in a search screen (see Figure 4). A search can be started for publications that match the search criteria. For instance, if the user types part of a keyword in the field Keyword and

TABLE 1 Example for Broader/Narrower Terms

Keyword	Broader terms	Narrower terms
Academic	Journal	–
Book	Publication	Manual, Introduction
Brochure	Publication	–
Industrial	Journal	–
Introduction	Book	–
Journal	Publication	Academic, Industrial
Manual	Book	Tutorial, Reference
Publication	–	Book, Brochure, Journal
Reference	Manual	–
Tutorial	Manual	–

selects the button Keyword Search, the application searches the multilingual terminology lexicon for keywords to match. The results can be shown in a drop down list (see Figure 5). Then a keyword from this list can be selected.

The multilingual terminology lexicon contains a list of key topics (i.e., keywords) in three languages, that is, Greek, English, and Spanish. A search for publications that match the search criteria is executed when the button Search is selected from the Search Screen. The number of publications found appears in the field Records Found. The search fields are as follows:

- *Title:* The title or part of the title of a publication;
- *Author:* The name or part of the name of an author;
- *Publishing year:* The issuing year of a publication;

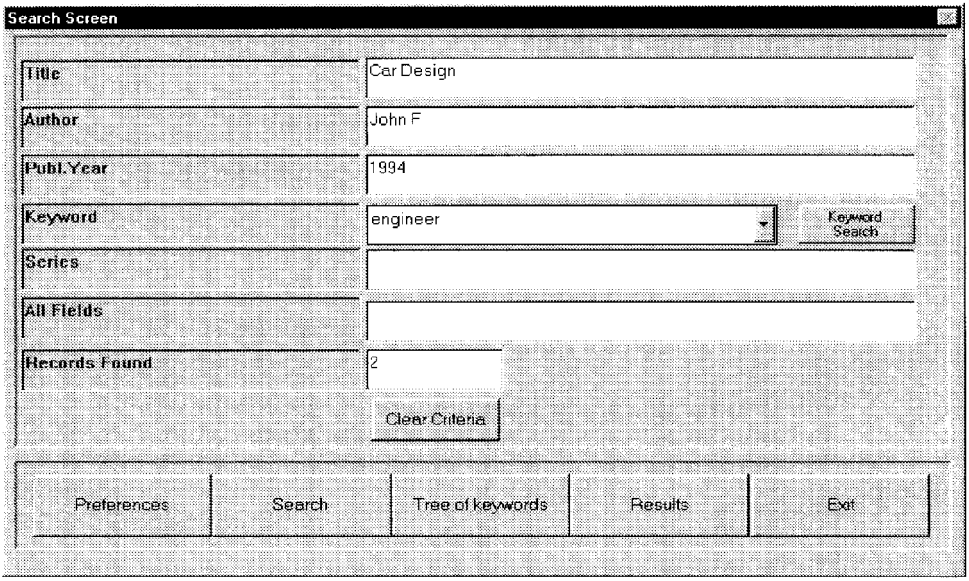


FIGURE 4. Search screen.

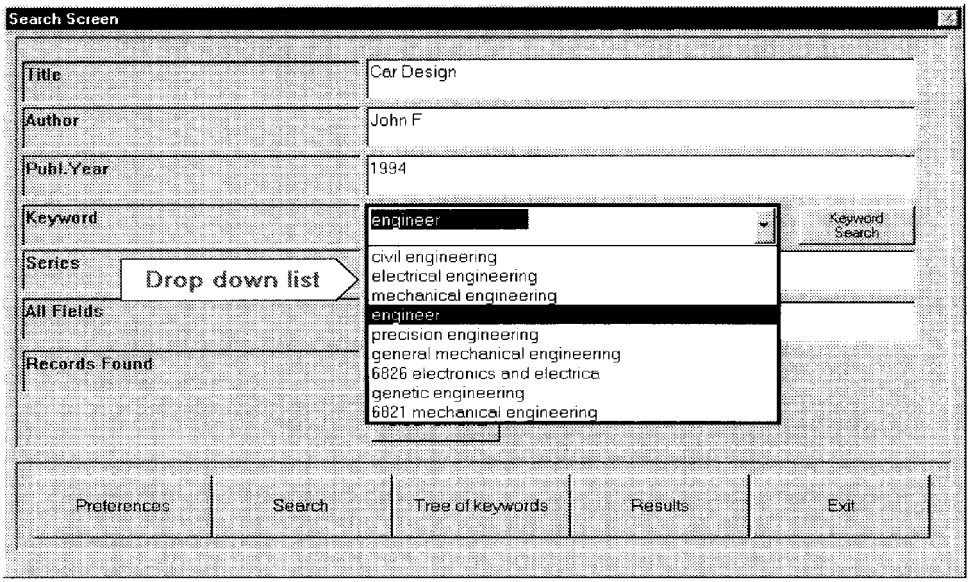


FIGURE 5. A list of keywords appear in a drop down list.

- *Keyword:* A keyword or part of a keyword;
- *Series:* The series or part of the series of a publication;
- *All fields:* All the fields associated with a publication (extended record).

When a search starts, all the fields associated with a publication are searched for a string that matches the string typed in this field.

The application searches for matching publications by using the "OR" operator between data entered in the above search fields. For example, if the search criteria contain the title "Democracy", the author "Cat Stevens," and the keyword "Ancient Greece," a search is performed for publications with the title "Democracy" or publications written by "Cat Stevens" or publications that contain the keyword "Ancient Greece." Finally, fields that are left blank are not included in the search criteria. For example, if the user leaves the fields Publishing Year, Keyword, Series, and All Fields blank, the search criteria include only the fields Title and Author.

Multilingual Presentation Tool

This tool allows localization of the user interface as well as presentation of the library database entries matching the user query in any of the supported languages. This can be achieved by translating publication titles and keywords, if needed. Particularly, the user has the capability of selecting the Message Language and the Output Languages (s)he prefers. More specifically, these languages stand for:

Input Language	Search Language(s)	Output Language(s)
English	English, Greek	English, Greek, Spanish

Search Criteria (keyword)	Search for	Results (Title)	(Keywords)
Democracy	*Democracy* *Δημοκρατία*	*Democracy in Ancient Greece* (English translation) *La Democracia en Grecia Antigua* (Spanish translation) *Η Δημοκρατία στην Αρχαία Ελλάδα* (original language)	*Democracy, Ancient Greece* *Democracia, Grecia Antigua* *Δημοκρατία, Αρχαία Ελλάδα*

FIGURE 6. A clarifying example.

- *Message language*: the language used for presenting labels and messages to the user as well as help screens;
- *Output language(s)*: one or more languages used for presenting the query results. The multilingual presentation tool utilizes:
- a *multilingual conversion table*, converting labels and messages into the Message Language, and
- a *simplified translation tool*, supporting the translation of book and journal titles and/or keywords into the specified Output Language(s).

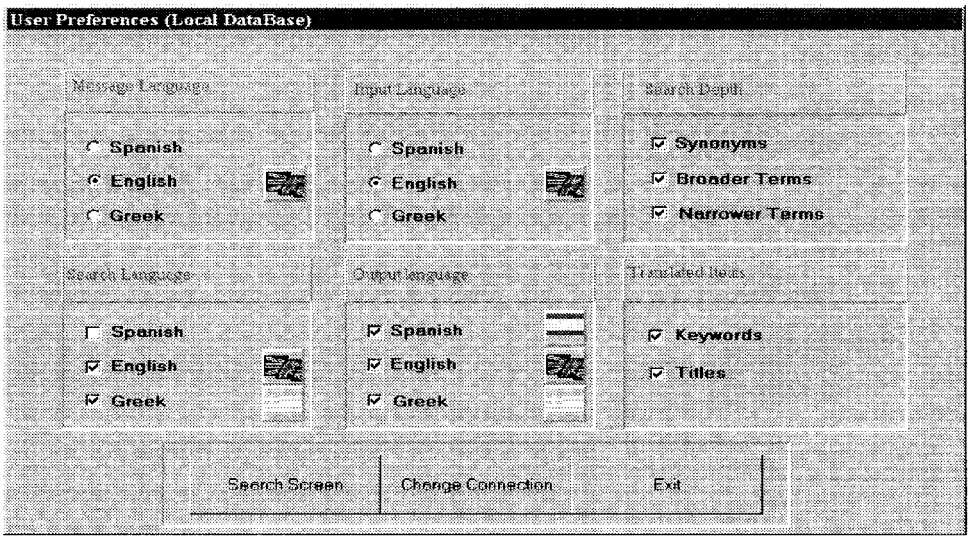


FIGURE 7. The user's preferences screen.

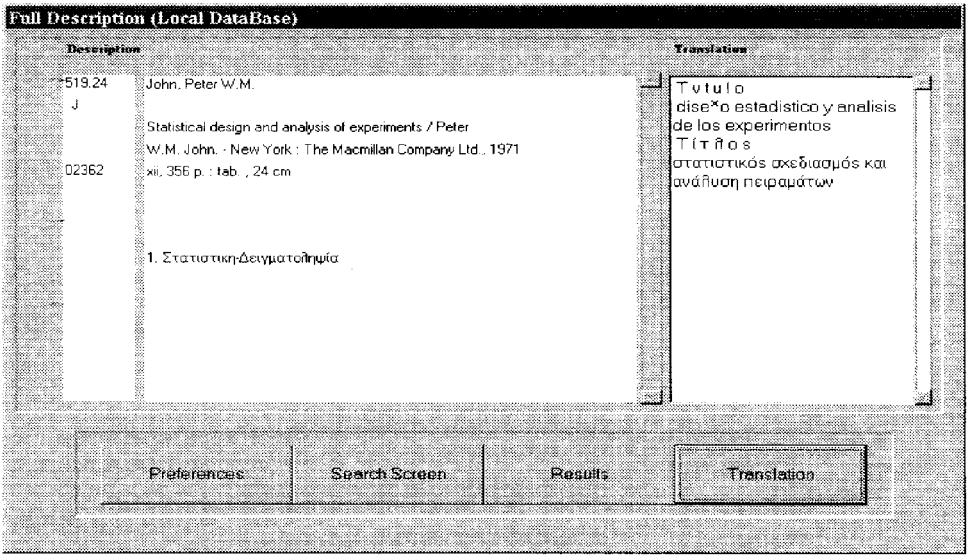


FIGURE 8. Typical full description of a retrieved book entry.

A Clarifying Example

The following clarifying example aims at further illustrating the above definitions. Let us assume that the user selects English as the Input Language, English and Greek as the Search Language(s), English, Greek, and Spanish as the Output Language(s), and the search criteria contain the keyword democracy. The multilingual search tool searches the library database for entries containing the keywords democracy (in English) and $\Delta\eta\mu\omega\kappa$–$\rho\alpha\tau\acute{\iota}\alpha$ (in Greek), and the multilingual presentation tool lists the publications that match these keywords and translates their titles and keywords into Spanish as is shown in Figure 6. Figure 7 shows the screen where the user enters his/her choices for this search.

Figure 8 shows a typical full description of another retrieved book entry that also contains its translation from English into Greek and Spanish. The system can also present the query results by using a short description of the retrieved book entry (i.e., including only the author(s), language of the title, title itself, and year of publication). At times a query can retrieve multiple records. In this case, the records are presented to the user ranked in alphabetical order. The original language of these records is also clearly labeled.

MULTILINGUAL RESOURCES

TRANSLIB is a system built through the integration of both new and existing information tools. Since the advantages of the use of reusable

resources are well known (Heid & McNaught, 1991), we tried to utilize general purpose and state-of-the-art resources where possible, in order to avoid the excessive cost of building a resource from scratch, as well as to ensure the quality and the adequacy of the resources. The following sub-sections describe in detail the resources of the presented system.

Multilingual Thesaurus

After an extensive market survey and taking into account the advantages and disadvantages of using an existing multilingual thesaurus, we decided that EUROVOC suits the needs of the TRANSLIB system. EUROVOC is based on ISO-5964/85, which is an international standard for the development of multilingual thesauri (Hradilova, 1995). EUROVOC is a multilingual thesaurus, which was originally built specifically for processing documentation of the European commission. It is implemented in the official languages of the European community. The edition we used was compiled in 1994 in nine languages (i.e., Danish, Dutch, English, French, German, Greek, Italian, Portuguese, and Spanish). Our version of EUROVOC contains only the following languages, which cover the needs of the current TRANSLIB system: *English*, *Greek*, and *Spanish*. All these languages have equal status-each descriptor in one language necessarily matches a descriptor in each of the other languages.

At the specific level of descriptors and nondescriptors, the internal structure of EUROVOC depends on semantic relations:

- scope note (SN)
- microthesaurus relationship (MT)
- equivalence relationship (UF, USE)
- hierarchical relationship (BT, NT)
- associative relationship (RT).

The EUROVOC thesaurus covers all the subjects that are of importance to the activities of the European community institutions:

- politics, international relations, European communities, law, economics, trade
- finance, social questions, education and communications, science, environment
- business and competition, employment and working conditions, transport, energy
- agriculture, forestry and fisheries, agrifoodstuffs, production
- technology and research, industry, geography, international organizations.

We mainly focused on the use of the following subjects: *economics*, *trade*, *finance*, *education and communications*, *science*, *business and competition*, *production*, *technology and research*, *energy*, *industry*, without excluding the others.

Technically speaking, EUROVOC is a single file (of more than 27MB) that contains a significant number of terms (more than 66,000) relevant to the above subjects. EUROVOC's architecture consists of a number of microthesauri. Each one of them consists of a set of terms for a given topic. Each record in a microthesaurus has several fields indicating the relation between the records and the terms themselves in the nine languages of the European community. Each term is given in two forms (all capital letters, and with upper and lower case letters) for each language. Two terms can have a hierarchical relation, that is, one can be narrower or broader than the other. Each record in the EUROVOC file fits on a single line. The fields occupy fixed positions on the line (e.g., 1–4, 9–13, 33–TL + 7, TL + 8, etc.) and have concrete names (e.g., physical blocklength, header, directory, separator, text, etc.)

Our goal was to transfer the data from EUROVOC to a commercial database. A set of routines was thus developed for the compilation of the EUROVOC thesaurus into files with a special format that can be easily read by database procedures. Several files constitute the TRANSLIB multilingual thesaurus derived from EUROVOC. There are separate files for the terms of each language. The terms are stored in the same order in each of the three files, so that the files are matching. If a term has no equivalents in one or more languages, there is a blank line in the corresponding position. The line number of each term in the files is used as a unique code during the building of the tree, which is stored in another file and does not include the terms themselves.

Thus, the multilingual thesaurus of TRANSLIB has been designed as a simple ASCII table file which contains all the material that is included in the EUROVOC thesaurus. Software was written in order to be able to extract the terms and relevant terms from this huge data store quickly. B + trees algorithms have been used to make the extraction process as quick as possible.

In conclusion, the multilingual thesaurus helps the user retrieve bibliography related to the relevant keywords (s)he has entered, as its sophisticated search mechanism automatically builds the tree structure of the keywords in all the preferred search languages and sends a complex query to the database. In the case where the user does not use the thesaurus, he may not retrieve all the available bibliographic records, since the keyword he uses to describe his query may not be the same as that used by the librarian for the record classification. Even when the user enters a query which has no keywords in the multilingual thesaurus, it is possible to update it by means of

an editing tool that caters for a number of data entry fields, such as New Keyword, Related Keyword, Keyword Language, Kind of Relationship, etc.

Multilingual Terminology Lexicon

The result of another extensive market survey encouraged us to adopt EURODICAUTOM as a terminology tool. However, after we had completed the decoding of EURODICAUTOM, a problem emerged. The terms of the multilingual thesaurus (EUROVOC) were entirely incompatible with those of the multilingual terminology lexicon (EURODICAUTOM). This meant that the library database could not be searched using keywords from the multilingual terminology lexicon, nor by using synonyms, broader or narrower terms of it.

The solution decided on was to create a multilingual terminology lexicon from the terms of EUROVOC. This was analyzed into two sections– the informative part and the data retrieval part. The former mainly consists of a file containing all the terms that are used as keywords in all the TRANSLIB supported languages. This file is a three–column file (one column per language), which contains the same term in each line for all the supported languages. The latter uses a binary search algorithm to achieve maximum efficiency. In this case, it was necessary to create a sequence of files (one for each search language), which contain an upper case description of terms and a numeric label used to locate the term and its description in all languages.

The multilingual terminology lexicon helps the user choose one term in the preferred Input Language, and search the database for this term as well as for all the equivalent terms in all the preferred Search Languages. Again the user can check the new keywords to be included in the multilingual terminology lexicon by means of an editing tool that caters for a number of data entry fields, such as New Keyword, English Keyword, Spanish Keyword, Greek Keyword, etc.

Multilingual Conversion Table

The multilingual conversion table is responsible for the interaction between the user and the system in the user's preferred language. It contains all the label fields, messages, and help screens of the user interface in all the supported languages and provides the user with the ability to both choose the language of the user interface and effectively change it at runtime according to his/her preferences. It is independent of any actual data store that is used as the base of the information storage. Finally, the information kept in the Multilingual Conversion Table is logically grouped into three files, one for each language. A new file can be used for a new supported

language, so that maximum source independence and management can be achieved.

The Multilingual Conversion Table is stored at the client's site in relational database tables. One table in the database includes the label fields, messages and on–line help information in the three languages supported by the project. Its internal structure is compatible with a future addition of equivalent terms in other European languages. The table has the following attributes:

- Code: a unique number that acts as the tuple's identifier
- Kind: reflects the type of the user interface item that is contained in the tuple
- EngContent: content in English
- SpaContent: content in Spanish
- GreContent: content in Greek.

When the user wants to customize any label fields, messages, or help screens, this can be accomplished through an editing tool that caters for a number of data entry fields, such as code label, Greek message, Spanish message, etc.

Library Database

The TRANSLIB system follows the UNIMARC standard so as to be easily portable to all library automation systems following the same standard. Regarding the library database, two medium–sized bibliographic databases have been utilized. The Greek one possesses about 10,000 titles concerning mainly engineering and computer science sectors, while the Spanish one possesses about 20,000 titles regarding political and financial topics. More specifically, the Greek library database contains 5,500 Greek titles, 3,500 English titles, and 1,000 titles in other European languages. On the contrary, the Spanish library database contains 12,000 Spanish titles, 6,300 English titles, and 1,700 titles in other European languages. Both databases were considered appropriate and representative to test the performance of the TRANSLIB system.

Simplified Translation Tool

The translation of the titles carried out by this tool should give the user an idea of the content of a retrieved book/journal. Even when the translation is not completely correct, it should help the user understand the subject of the document (e.g., consider the translation "Democracy in *the* Greece").

Translating titles is, in general, easier than translating a whole sentence or even entire texts. The simplest title is a single noun while the most complex one may be a whole phrase. Few titles, however, are composed of a long phrase, i.e., the combination of both noun and prepositional phrases. Furthermore, a title is selected carefully by the author to represent the content of a book. This fact means that a title must not be ambiguous in order to be comprehensible immediately by the reader. In most cases, syntactic or referential ambiguities can easily be solved. Additionally, titles are usually typed carefully by the editors of a book, so that no grammatical or syntactic errors occur. Needless to say, the titles of books and journals are totally different from those found in a newspaper (i.e., news headlines).

Looking at the translation of the titles from a computational point–of–view, the translation process can be facilitated further by making a set of assumptions. These can be described as follows:

- Translation from one language into another is realized via English which is used as an *interlingua*.
- For each title only one translation is provided. In addition, a set of titles may have the same translation (i.e., many–to–one translation).
- Only the main title of a bibliographic entry is translated. Subtitles, version numbers, etc. are ignored.
- The titles are considered to be linguistically correct. Hence, syntactically incorrect interpretations are ignored.
- No semantic and/or pragmatic information is used. All the titles are referred to the scientific domain of engineering and computer science. Moreover, for every word in the dictionary there is only one translation.
- Finally, the input as well as the output files of the translators are ASCII files containing one title per line. If a title cannot be translated, the output file contains an empty line for it.

The selection of an interlingua was adopted to achieve system robustness and reduce the translation costs. Translation between a pair of languages implies translation to and from interlingua. Additionally, the upgrade of the system with a new supported language would only require the implementation of a lingware for the new language, able to provide translation to and from interlingua. We selected English as the interlingua, since it was found that it is the most widespread foreign language among the OPAC users.

In fact, our interlingua is not plain English but an artificial language, that is, an abstract form that represents the structure of the title based on the English translations of its consisting words. Thus, the interlingua is composed of a list of structures that correspond to the words of the title. These structures have the following form in PROLOG language for a given word of the line:

```
word(English_Lemma,Part_of_Speech,Number,Case)
```

where English_Lemma is the English translation of the Greek or Spanish word (in singular number) of the title, and Part_of_Speech, Number, and Case are the corresponding morphological information of this word in the source language (i.e., Greek or Spanish). For instance, the interlingua for the Greek title "*Χραμμική άλγεβρα και αναλττική γεωμετρία*" ("Linear algebra and analytical geometry") would be the following list:

```
[word(linear,adjective,singular,nominative)
 word(algebra,noun,singular,nominative)
   word(and,conjunction,_._)
   word(analytical,adjective,singular,nominative)
 word(geometry,noun,singular,nominative)].
```

Taking all the above points into account and in order to strengthen our claim that an AI–based translation tool is needed in the translation process, we tried *two translation scenarios*. We first implemented a simple word–by–word translation (i.e., without the use of any AI tool), in order to simplify the translation process and consequently minimize the computational cost. The evaluation of this approach* has shown that only a small percentage of the translated titles (about 25%) were comprehensible, particularly in trans–lations from Greek to English, Spanish to English, and vice versa (due to interlingua problems in this case).

We decided, then, to implement a more sophisticated approach that would improve the quality of the translations by taking advantage of AI–based methods (e.g., use of natural language processing tools). For instance, the following tools for the implementation of the Greek and English ling–ware (i.e., the tools that provide translations from Greek to English and vice versa) were used:

- An existing two–level morphological analyzer for morphological recogni–tion and synthesis of Modern Greek words (Antworth, 1990; Sgarbas et al., 1995). This is a PC KIMMO–based analyzer that was designed for general purpose morphological analysis of Modern Greek and uses a 30,000–word lexicon. According to the two–level morphological model, the surface form (i.e., the linguistically correct form) of a word is derived by applying a set of rules to the lexical form (i.e., the sequence of morphemes) of that word and vice versa. The main advantage of such a model is that these rules are applied in parallel and in both directions, that is, surface to lexical form (i.e., morphological analysis) and lexical to surface form (i.e.,

* We selected randomly 1,000 Greek, Spanish, and English titles and applied to them the first translation scenario. The results (i.e., the translated titles of the TRANSLIB output) were compared with the translations by native speakers in Greek, Spanish, and English.

morphological synthesis). In linguistic terms, the two–level rules are not processing rules, but more like the realization rules of stratification linguistics.
- A public domain two–level morphological analyzer (similar to the previous one) for morphological recognition and synthesis of English words (Antworth, 1990). This is a PC KIMMO–based analyzer that was designed for general purpose morphological analysis of English and uses a 10,000–word lexicon.
- Two syntactic parsers (i.e., one for modern Greek and one for English) built from scratch and able to recognize the basic phrase structures of the title. These parsers are based on a context–free grammar model consisting of 28 rewriting rules for each one of them. As an example, the four first rules of the modern Greek grammar are given below:

```
Title⇐Phrase.
Title⇐Phrase, Conjunction, Phrase.
Phrase⇐NounPhrases, Prepositional Phrases.
Phrase⇐Prepositional Phrases.
```

By applying these rules to an input title, a form representing its syntax is produced. For instance, for the aforementioned Greek title "*Χραμμική άλγεβρα και αναλττική γεωμετρία*" ("Linear algebra and analytical geometry"), the output of the English parser in PROLOG language would be as follows:

```
title([nounphrase([adjective(linear),noun(algebra)]),conjunction(and),
nounphrase([adjective(analytical),noun(geometry)])]).
```

- Two medium–sized bilingual (i.e., Greek–English and English–Greek) morphological dictionaries (about 5,000 lemmas) containing words from the scientific domains of engineering and computer science found in the Central Library of the University of Patras. In particular, these dictionaries stemmed from the PC KIMMO–based dictionaries and translation information as well as new words from the previous scientific domains were added.

Similar tools for the implementation of the Spanish lingware were utilized. Just to illustrate the translation process from a Greek title into a corresponding English one, we give below the sequence of transitions among the different tools:

i. Greek Title → Greek Morphological Analyzer → Morphologically Analyzer Greek Title

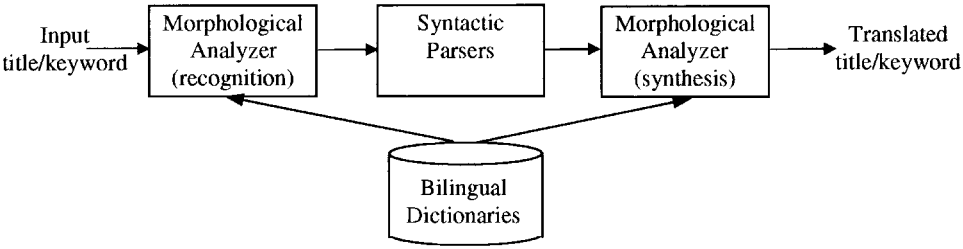


FIGURE 9. Outline of the translation tool.

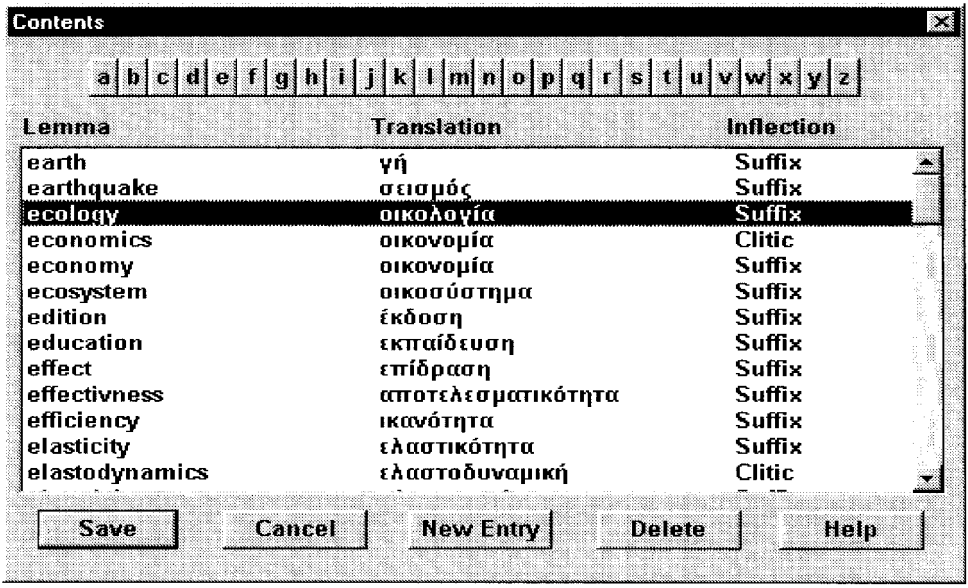


FIGURE 10. Editing tool for the bilingual dictionaries.

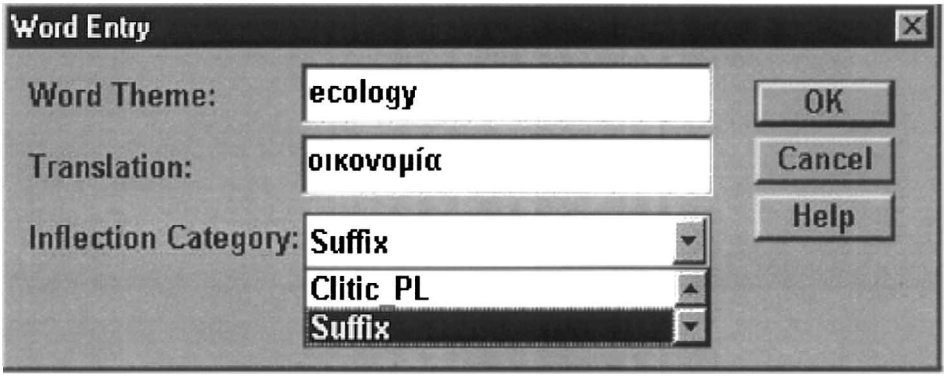


FIGURE 11. Inserting a new entry.

ii. Morphologically Analyzed Greek Title → Greek Syntactic Parser → Abstract Syntactical Form
iii. Abstract Syntactical Form → Greek–English Translator, Greek–English Bilingual Dictionary → Translated Abstract Form
iv. Translated Abstract Form → English Morphological Analyzer → Morphologically Correct English Title
v. Morphologically Correct English Title → English Syntactic Parser → Well–Formed English Title

An outline of the simplified translation tool is shown in Figure 9. In contrast to the word–by–word translation, approximately 86% of the translated titles are comprehensible.* It has to be noted here that the quality and comprehension of the provided translations depend on the scientific domain of the library database. For instance, the Greek word "σειρά" has two meanings: (i) series (that is mainly used in mathematical context), and (ii) line (that has a general use). Hence, if the library database comprises books/journals concerning the scientific domain of mathematics, then the former translation should be included in the bilingual dictionaries.

Finally, an editing tool is used for updating entries in the English–Greek and Greek–English dictionaries. The user is able to update the nouns and adjectives (i.e., the inflectional parts of speech) that are included in the selected bilingual dictionary (see Figure 10). A new entry may be inserted or an existing one may be changed or deleted. When inserting a new entry, the user must insert the new lemma as well as its translation. Additionally, (s)he should also select an inflectional category in order to determine the various word forms of the new entry (see Figure 11).

EVALUATION

The TRANSLIB system has been tested in the library databases in both Greece and Spain under real life conditions. Two parameters were evaluated: (i) the subject search, i.e., the proportion of citations found by the user referring to a given subject, and (ii) the user satisfaction, i.e., the impression of the user with respect to the quality of the service provided.

The first parameter was measured by means of the standard 11–point average precision for our DICT method. The 11–point average precision is a conventional performance measure for evaluation of text retrieval systems (Salton, 1989). Given a query, for recall thresholds of 0%, 10%, 20%, ...,

* Again, we used the same 1,000 Greek, Spanish, and English titles and applied to them the second translation scenario. The results (that is, the translated titles of the TRANSLIB output) were compared with the translations by native speakers in Greek, Spanish, and English.

100%, the system assigns in decreasing score order as many titles as needed until a given recall level is achieved, and computes the precision value at that point. Then, the precision values of individual queries are averaged over all test queries. These averaged precision values can be further averaged over the recall thresholds, to obtain a single–numbered global measure. The resulting global average is referred to as the 11–point average precision.

We carried out the measurement of the first parameter in the *Greek library database*, initially with the previous OPAC system (monolingual retrieval) and then with the TRANSLIB system (multilingual retrieval). We used *1,000 test queries* from the scientific domain of engineering and com–puter science. The input language for these queries was either Greek or English, and the search criteria comprised either a unique search field or a combination of two or more search fields. The vast majority of these queries (about 85%) consisted of isolated words or short phrases. Only 15% of the above queries formed long phrases. The recall–precision values for these two systems are shown in Table 2. The DICT approach proves efficient enough for queries consisting of isolated words or at best short phrases.

The second parameter was evaluated by means of a short questionnaire which was used to gather the general impression on the system and its fea–tures. This questionnaire comprised 27 questions, which were classified in four distinct parts: (i) personal information regarding the evaluator, (ii) results on the use of TRANSLIB, (iii) personal opinion of TRANSLIB, and (iv) comments. In some cases, the questionnaire reply boxes ranged from 1 to 5 by degree of satisfaction. Finally, a total of 42 library users filled the ques–tionnaires. They were selected on the basis of the following criteria: (a) to be familiar with OPACs, (b) to possess a good knowledge of English and/or Spanish, (c) to have different sex and age, and (d) to be willing to participate in the evaluation process.

TABLE 2 Recall–Precision Values for the DICT Approach

Monolingual Retrieval		Multilingual Retrieval	
Recall	Precision	Recall	Precision
0.0	0.90	0.0	0.63
0.1	0.86	0.1	0.50
0.2	0.83	0.2	0.45
0.3	0.80	0.3	0.39
0.4	0.77	0.4	0.34
0.5	0.72	0.5	0.30
0.6	0.68	0.6	0.23
0.7	0.59	0.7	0.21
0.8	0.52	0.8	0.18
0.9	0.50	0.9	0.12
1.0	0.35	1.0	0.10

The results of the evaluation are encouraging and pinpoint the strength as well as the potential weaknesses of this system. Some of the most important results are listed below:

Multilingual Search

- 95% of the users performed multilingual searches in the library database rather than monolingual ones;
- approximately 45% of the users felt rather satisfied or completely satisfied about the results of their search, while 9% of them were dissatisfied;
- less than 25% of the dissatisfied users considered multilingual aspects not supported in TRANSLIB responsible for their dissatisfaction;
- approximately 85% of the users considered the search via the multilingual thesaurus more accurate and complete

Translation of Titles

- approximately 80% of the users considered the translation of titles rather useful or very useful;
- over 55% of the users considered the translations of titles comprehensible. Note that users with high familiarity with OPACs tended to be slightly more satisfied with the quality of the translation;
- over 90% of the users found the functionality of the translation component very easy.

Multilingual Presentation

- over 70% of the users considered the system easy to use by anyone, not just experts;
- over 60% of the users considered TRANSLIB more friendly, useful, and easier to understand than other OPACs (with no multilingual features) they had used in the past. It is noteworthy that users with a high familiarity in OPACs were considerably more satisfied;
- finally, over 65% of the users found the help and error messages of TRANSLIB clear, useful, and adequate.

As noted earlier, the selection of an interlingua affects the quality as well as the comprehension of the translation of titles. Hence, translations to and from English are more comprehensible than from Greek to Spanish and vice versa. Nevertheless, approximately 95% of the performed translations are translations to and from English, since the vast majority of the book/journal

titles contained in a library consists of titles in the local language and English. The satisfaction of the users with the multilingual thesaurus and the overall search results they obtained from the system in contrast to other OPACs was remarkable. The localization of the user interface was one of the main reasons why the users favored this system. In conclusion, the reaction of the users to the TRANSLIB system appears to be closely dependent on their previous experience.

- Users find multilingual search very useful.
- Users with a higher familiarity with OPACs find TRANSLIB a simple, friendly, and easy to learn system. Users with a lower familiarity with OPACs find it more difficult to use TRANSLIB. This is a satisfactory conclusion, as TRANSLIB neither intends to replace OPAC nor makes its baseline operation simpler. It is simply intended to overcome the linguistic barriers posed by multilingual library holdings.
- Users appear to search for titles in languages in which they are familiar. For this reason they tend to show limited appreciation of the usefulness of the translation of titles, while they expect a better translation.
- Although the user reaction is positive, there appears to be room for improvement with respect to the terminology used, the help and error messages, and the way that the system keeps the user informed of what is being done at any time.
- The satisfaction of the users (with respect to their search results) depends highly on both the TRANSLIB system as well as on the bibliographic database and the library holdings. As a result it is difficult to assess the impact of TRANSLIB on the quality of the search results. More extensive experiments would be needed for this purpose, preferably in a distributed, transnational bibliographic database.

CONCLUSIONS

As we described earlier, despite the fact that there are several OPACs supporting the automation of retrieving bibliographic information, only a few of them support some kind of bilingual or even multilingual access to libraries catalogues (McAllistair, 1987; Hug & Noethinger, 1988; Slater, 1991; Butcher, 1993). In this paper we presented a system that aims at solving this problem. TRANSLIB allows multilingual search as well as multilingual presentation of the query results and consists of an integration of state-of-the-art existing information tools and tools built from scratch. The integration of these tools was a difficult task, since the former were built for different purposes and, some of them, for different operating systems. The multilingual functionalities of TRANSLIB help the user, to a high degree,

improve significantly the results of a query and allow him/her to select the presentation of the results in the preferred language.

For all these reasons, we feel that libraries (e.g., public, private, university, etc.), book trade enterprises, appropriate consultancy services and library schools can directly benefit from a system of this type. TRANSLIB should have a remarkable impact on the improvement of library services as well as on the easy access to the wealth of knowledge held in libraries. The results of its performance on the Greek library database and its pilot installation in the Central Library of University of Patras strongly encourage us to believe that it will have a successful commercial exploitation in the near future.

Our short–term goal is to endow TRANSLIB with the capability to support bibliographic information retrieval from remote databases through INTERNET servers. In addition, we plan to further improve the present system by adding new languages in the existing framework. Our long–term goal is to enrich it with new AI tools. In particular, we intend to provide the translation of entire abstracts rather than just the titles and improve the accuracy of the search results based on intelligent information extraction.

REFERENCES

Antworth, E. 1990. PC–KIMMO: A two–level processor for morphological analysis. Summer Institute of Linguistics. Dallas, Texas.

ANSI/NISO Z39.50. 1995. Information retrieval (Z39.50). Application service definition and protocol specification.

Buchinski, E. J., W. L. Newman, and M. J. Dunn. 1976. The automated authority subsystem at the National Library of Canada. *Libr. Autom.* 9(4):279–298.

Butcher, R. 1993. Multi–lingual OPAC developments in the British Library. *Program* 27(2):165–171.

Cousins, S. A., and R. J. Hartley. 1994. Towards multilingual online public access catalogues. *Libr. Autom.* 44(1):47–62.

Fluhr, C., P. Ortet, F. Elkateb, K. Gurtner, and V. Semenova. 1996. Distributed multilingual information retrieval. In *Proc. Multilinguality in Software Industry: The AI Contribution* (*MULSAIC'96*) *Workshop*. Budapest, Hungary.

Gibb, F. 1993. Knowledge–based indexing in SIMPR: Integration of natural language processing and principles of subject analysis in an automated indexing system. *Docum. Text Manag.* 1(2):113–125.

Gibb, F., and G. Smart. 1991. Knowledge–based indexing: The view from SIMPR. In *Libraries and Expert Systems*, ed. C. MacDonald and J. Weckert, 38–48. London: Taylor Graham.

Gilarranz, J., J. Gonzalo, and F. Verdejo. 1997. Language–independent text retrieval with the EuroWord–Net multilingual semantic database. In *Proc. MULSAIC'97 Workshop*, ed. C.D. Spyropoulos, 9–16. August 25, Nagoya, Japan.

Heid, U., and J. McNaught. 1991. Eurotra–7 study: Feasibility and project definition study on the reusability of lexical and terminological resources in computerised applications. Final Report. Brussels, Belgium.

Hradilova, J. 1995. Thesaurus EUROVOC– Indexing language of the European Union. *Infoc.* 1(3):66–69.

Hug, H., and R. Noethinger. 1988. ETHICS: An online public access catalogue at ETH–Bibliothek. *Program.* 11(2):133–142.

Kikui, G., Y. Hayashi, and S. Suzaki. 1996. Cross–lingual information retrieval on the WWW. In *Proc. MULSAIC'96 Workshop*. Budapest, Hungary.

Leeves, J. 1994. Library systems in Europe: A directory & guide. London, England: TFPL.

Lehtola, A., and T. Honkela. 1996. Multilinguality in electronic commerce– Research issues. In *Proc. MULSAIC'96 Workshop*. Budapest, Hungary.

McAllistair, C. 1987. The online public access catalogue in DOBIS–LIBIS. *Program*. 21(1):25–36.

Morris, A. 1992. The Application of Expert Systems in Libraries and Information Centres. London: Bowker–Saur.

Oard, D. 1997. Alternative approaches for cross–language text retrieval. In AAAI Spring Sympositum on Cross–Language Text and Speech Retrieval. March 24–26, Stanford University, Stanford, CA.

Pasanen–Tuomainen, I. 1992a. Analysis of subject searching in the TENTTU books database. In *Proc. International Association of Technological University Libraries (IATUL)* 1:72–77. Cambridge, Mass.

Pasanen–Tuomainen, I. 1992b. Monitoring online catalogues in the Nordic Technological University libraries. *Nordinfo Nytt*. 4:23–27.

Salton, G. 1989. Automatic text processing: The transformation, analysis, and retrieval of information by computer. Pennsylvania: Addison–Wesley.

Schütz, J. 1996. Intelligent web–based information services. In *Proc. Multilinguality in Software Industry: The AI Contribution (MULSAIC'96) Workshop*. Budapest, Hungary.

Sgarbas, K., N. Fakotakis, and G. Kokkinakis. 1995. A PC–KIMMO–based morphological description of modern Greek. *Liter. Ling. Comp*. 10(3):189–201.

Slater, R. 1991. Authority control in a bilingual OPAC: MultiLIS at Laurentian. *Library Resources Technical Services* 35(4):422–458.

Stamatatos, E., S. Michos, K. Patelodimou, and N. Fakotakis. 1997. TRANSLIB: An advanced tool for supporting multilingual access to library catalogues. In *Proc. MULSAIC'97 Workshop*, ed. C.D. Spyropoulos, 9–16. August 25, Nagoya, Japan.

Synellis, C. 1995. TRANSLIB: User survey, Report 1.1, Patras, Greece.

Yang, Y., R. Brown, R. Frederking, J. Carbonel, Y. Geng, and D. Lee. 1997. Bilingual–corpus based approaches to translingual information retrieval. In *Proc. MULSAIC'97 Workshop*, ed. C.D. Spyro–poulos, 9–16. August 25, Nagoya, Japan.

Zhang, X., J. Liu, and Atwell 1997. A multilingual information retrieval tool hierarchy for a WWW virtual corpus. In *Proc. MULSAIC'97 Workshop*, ed. C.D. Spyropoulos, 9–16. August 25, Nagoya, Japan.