The 2nd Workshop on Multilinguality in Software Industry

- MULSAIC'97 -

Costas Spyropoulos

W21

IJCAI-97

15th International Joint Conference on Artificial Intelligence

August 23-29,1997
Nagoya Congress Center, Nagoya, Japan

Sponsored by:The International Joint Conferences on Artificial Intelligence, Inc.
The Japanese Society for Artificial Intelligence (JSAI)

TRANSLIB: An Advanced Tool for Supporting Multilingual Access to Library Catalogues

E. Stamatatos*, S. Michos*, C. Patelodimou*, and N. Fakotakis**

* KNOWLEDGE S.A.
N.E.O. Patron-Athinon 37
264 41 Patras, GREECE
stamatatos@wcl.ee.upatras.gr

** University of Patras
Wire Communications Laboratory

Abstract

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 stemmed from the integration of both new and already existing advanced multilingual information tools. By making use of some AI-based methods this system takes full advantage of these resources in order to provide multilingual access to library catalogues. Among its striking features, it enables searching in multiple languages, multilingual presentation of the query results, and localization of the user interface. TRANSLIB has been currently tested in existing medium-sized bibliographic databases. Early 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.

1 Introduction

Present libraries have automated their trivial 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, 94]. 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, 92]. 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 already improved the abilities of several systems [Gibb & Smart, 91; Gibb, 93].

Nevertheless, libraries did not pay special attention to multilingual features of OPACs, despite the fact that it is a serious problem [Cousins, 94]. The National Libraty of Canada was one of the first libraries that offered multilingual access to its users, even in the form of controlled bilingual (i.e., English/French) authority files [Buchinski *et al.*, 76]. The ETHICS project produced an OPAC with multilingual user interface and help screens as well as a subject index in three languages (i.e., French, German, and English) [Hug & Noethinger, 88].

So far, multilinguality in OPACs was dictated by the needs of a specific multilingual community and was restricted to the provision of bilingual or multilingual lists of controlled terms such as authority control files, subject headings, and thesauri [McAllistair, 87; Slater, 91; Butcher, 93]. Multilinguality as a potential problem for the common user of the library was only explicitly investigated in the NORDINFO survey [Pasanen-Tuomainen, 92a; Pasanen-Tuomainen, 92b]. According to the results of this survey, multilingual access to online catalogues may 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 Spanish Agency of International Cooperation, and the Municipal Library of Patras [Synellis, 95]. These surveys were aimed at the analysis of the attitude of the users towards the OPAC they used and the investigation about their eventual need of a multilingual tool. The results of these surveys are essentially independent of the sex, educational status, familiarity of the computer use, and frequency of the OPAC use of the sample users. The analysis of the results was very illuminating: approximately 75% of the users were whether moderate or not satisfied about the results of their searches in the OPAC. Over 80% of the unsatisfied users noted that the hits were not in the user's native language or that the search was based only on a single language. Furthermore, approximately 70% of the users were interested in bibliography written in foreign languages, and 65% of them who use books in only one language consider searching in multiple languages whether useful or very useful.

On the other hand, despite the penetration of AI to 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.*, 96; Kikui *et al.*, 96]. Furthermore, the previous surveys have also showed that 64% of the users consider a potential translated title presentation in their native language very useful whether they use one or more languages [Synellis, 95].

TRANSLIB is a system that 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 already existing advanced information tools and has been tested in existing medium-sized bibliographic databases in Greece and Spain. The aforementioned tools are characterized by 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 future prospects are given towards making TRANSLIB a marketable product.

2 Overview of TRANSLIB

TRANSLIB is an OPAC that provides the users with the capability of multilingual access to library catalogues. TRANSLIB is fully-implemented, runs under Windows 95, and currently supports three languages, that is English, Greek and Spanish. These three languages are considered to be sufficient for both demonstrating the feasibility of this endeavor and pinpointing potential problems in the envisioned adoption of additional European or non-European languages. An outline of the presented system is depicted in Figure 1.

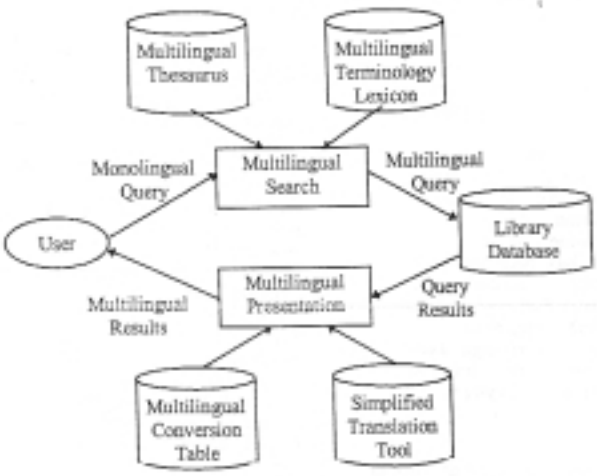

Fig. 1: An outline of TRANSLIB system

TRANSLIB supports 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 that specifies a client/server based protocol for information retrieval that was originally proposed for use with bibliographic information [ANSI/NISO Z39.50, 95].

In more detail, TRANSLIB comprises two basic tools, each one of significant self interest:

(i) the **Multilingual Search Tool** which supports the retrieval of multilingual bibliographic information, and

(ii) The **Multilingual Presentation Tool** which allows multilingual presentation of the retrieved information.

The following subsections illustrate the features of the above tools. By taking advantage of them, TRANSLIB offers the following facilities in contrast to up-to-date library systems [Leeves, 94]:

- Localization of user-interface (i.e., English, Greek, or Spanish) is supported.
- The user is able to give 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 searching which can be based on narrower and/or broader terms of the query.
- Translation of the retrieved books/journals titles and keywords in any language is provided in

order to give a sense about the content of these books/journals to the user.

2.1 Multilingual Search Tool

This tool enables the user to enter the search query in the language (s)he prefers as well as to select the languages in which the matching entries have to be found. Essentially, this tool performs a conversion from monolingual to multilingual queries. Particularly, the user has the ability to select the input language and the search languages (s)he wishes for. More concretely, these languages stand for:

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

Furthermore, the user is able to determine the search depth of one's search by selecting to search with synonyms, and/or narrower terms, and/or broader terms of one's 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 the use of synonyms, narrower and broader terms.

2.2 Multilingual Presentation Tool

This tool allows localization of the user interface as well as the library database entries that match the user query to be presented in anyone of the supported languages 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 wishes for. More concretely, these languages stand for:

- **Message language:** the language used for presenting labels and messages to the user as well as for 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
- **Simplified Translation Tools**, that support the translation of book and journal titles and/or keywords into the specified output language(s).

2.3 A Clarifying Example

The following clarifying example aims at further illustrating the above definitions. Consider that the user selects English as the *input language*, English and Greek as the *search languages*, English and Spanish as the *output languages* and the *search criteria* contain the keyword *Democracy*. The Multilingual Search Tool searches the library database for entries that contain the keywords *Democracy* (i.e., in English) and *Δημοκρατία* (i.e., in Greek), and the Multilingual Presentation Tool gives the publications that match these keywords and translates their titles and keywords in English and Spanish as it is depicted in Figure 2.

Input Language	**Search Language**	**Output Language**
English	English, Greek	English, 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*

Fig. 2: A clarifying example.

Figure 3 shows the screen where the user chooses his/her preferences and Figure 4 shows the full description of a retrieved book entry (containing its translation from English into both Greek and Spanish).

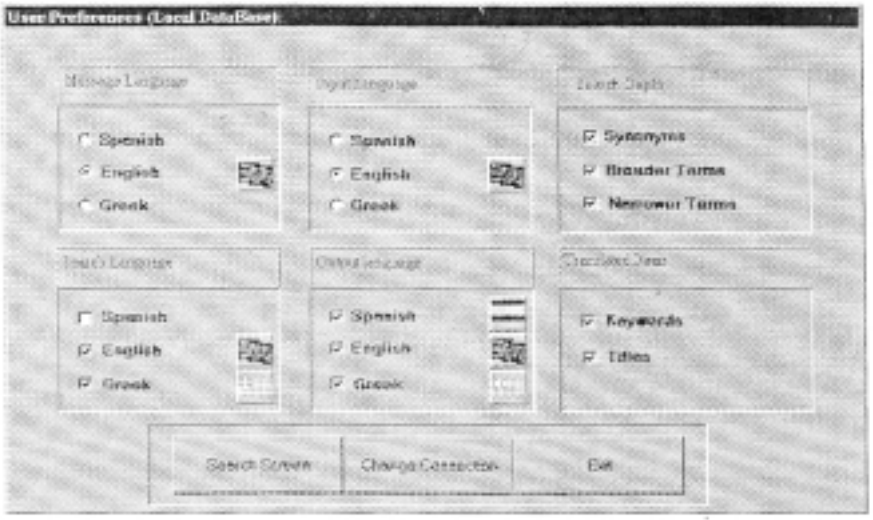

Fig. 3: The user's preferences screen.

Multilingual Resources

ANSLIB is a system built by the integration of new and already existing advanced information s. Since the advantages of the use of reusable urces are well-known [Heid & McNaught, 91], tried to utilize general-purpose and state-of-the-art resources, where it was 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 subsections describe in detail the resources of the presented system.

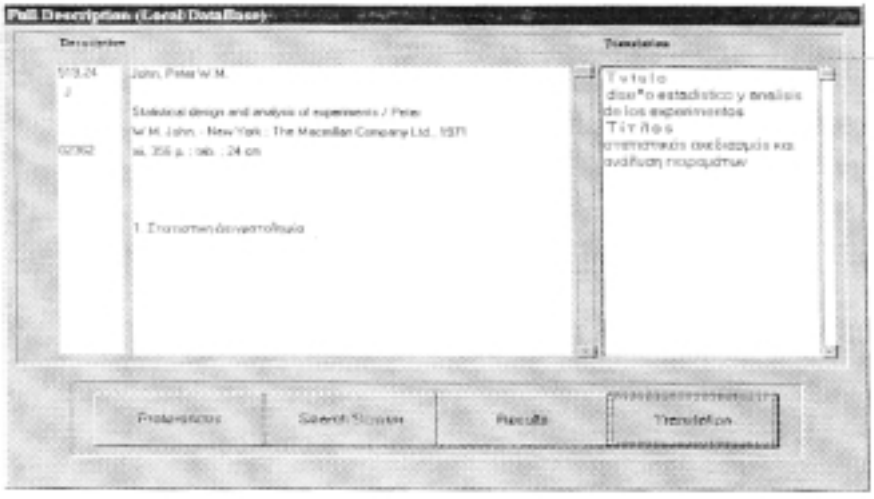

Fig. 4: Full description of a retrieved book entry.

3.1 Multilingual Thesaurus

After an extended market research 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 which is an international standard for the development of multilingual thesauri [Hradilova, 95]. EUROVOC is a multilingual thesaurus which was originally built up especially for processing the documentary information of the European Community institutions. 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).

The files that constitute the TRANSLIB Multilingual Thesaurus derived by EUROVOC have the following format. There is a file for the terms of each language, that is, one file with the Greek terms, one with the English terms, and one with the Spanish ones. The terms are stored with the same order in the three files so as to be equivalent with each other. If a term has no equivalents in one or more languages, there is a blank line in the specific 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.

In conclusion, the Multilingual Thesaurus helps the user retrieve bibliography about all the relevant keywords he wants, as its sophisticated search mechanism builds automatically the tree structure of the keywords in all the preferred search languages and it sends a complex query to the database without the participation of the user himself. In case 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 with which the librarian used for the record classification.

3.2 Multilingual Terminology Lexicon

The Multilingual Terminology Lexicon contains the same keywords in all the supported languages. The Multilingual Terminology Lexicon is a subset of the Multilingual Thesaurus. Both tools must contain the same keywords in order to be fully compatible. It is implemented by the three files for Greek, English and Spanish terms. The tool facilitates the user to choose one term in the preferred input language and search the database for this term as well as for the equivalent terms in all the preferred search languages.

3.3 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 choose the language of the TRANSLIB user interface and effectively change it at runtime according to his/her preferences. The Multilingual Conversion Table is implemented as a client part, and 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.

3.4 Library Database

The TRANSLIB system follows the UNIMARC standard so as to be easily transportable 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 the engineering and informatics sectors while the Spanish one possesses about 20,000 titles regarding the political and financial sectors. Both databases were considered appropriate and representative for the successful performance of the TRANSLIB system.

3.5 Simplified Translation Tool

The titles' translations derived by this tool have to give the user a sense about the content of the retrieved book/journal. In cases, however, that the derived translation is not completely correct but it manages to help the user understand what the book is dealt with, this translation is considered to be comprehensible (e.g., consider the translation "*Democracy in the Greece*"). In more detail, the book/journal titles are translated taking into account the following limitations:

- 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).
- Finally, only the main title of a bibliographic entry is translated. Subtitles, version numbers, etc. are ignored.

The selection of an interlingua was made for achieving system robustness and reducing the translation cost. Hence, in this case, translation between a pair of languages requires only translation to and from interlingua. Additionally, the upgrade of the system with a new supported language would only require the implementation of a lingware tool for the new language able to provide translation to and from interlingua. Moreover, we selected English as the interlingua since it was found that it is the most widely spread foreign language among the OPAC users. Therefore, it seems reasonable the vast majority of translations to be consisted of translations from the user's native language to English and vice versa.

The simplest title is a single noun whilst the most complex one may be a whole sentence. The vast majority of titles, however, is composed of a complex phrase, that is, the combination of noun phrases and prepositional phrases. Furthermore,

titles are carefully selected by the author for representing the content of the book. This fact means that a title must not be ambiguous in order to be comprehensible instantly. In most cases, syntactic or referential ambiguities can be easily solved. Additionally, titles are usually carefully typed by the editors of the book, so there are no grammatical or syntactic errors.

Taking all the above points into account, we first implemented a very simple word-by-word translation (i.e., without the use of morphological analyzers and syntactic parsers) 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, especially in translations from Greek to Spanish and vice versa (due to interlingua problems in this case). Therefore, we decided 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).

In particular, we used the following tools for the implementation of the Greek lingware tool (i.e., the tool that provides translations from Greek to English and vice versa):

- an already existing two-level morphological analyzer able to be used for morphological recognition as well as morphological synthesis [Antworth, 90; Sgarbas et al., 95], and
- a syntactic parser built from scratch able to recognize the basic phrase structures of the title. This parser is based on a simple context-free grammar and is able to analyze correctly over the 90% of the tested titles.

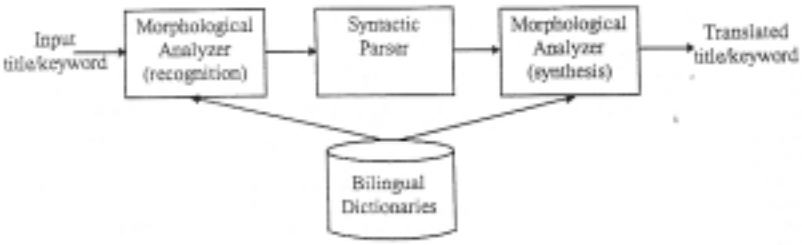

Fig. 5: Outline of the translation tool.

An outline of the Simplified Translation Tool is shown in Figure 5. This tool utilizes two medium-sized bilingual morphological dictionaries (about 5000 lemmas) containing words from the scientific domains of engineering and informatics found in the Central Library of the University of Patras. For each entry in these dictionaries there is only one translation. In this case, approximately 85% of the titles' translations are comprehensible. It has to be noted that the quality and the comprehension of the provided translations depends on the scientific domain of the library database. For instance, the Greek word '*σειρά*' has two meanings: (i) *series*, and (ii) *line*. If the library database comprises books/journals concerning the scientific domain of mathematics, then the former translation has to be selected.

4 Evaluation

The TRANSLIB system has been tested in the aforementioned library databases in both Greece and Spain. Two parameters have been evaluated: (i) the *subject search*, that is, the proportion of citations found by the user referring to a given subject, and (ii) the *user satisfaction*, that is, the impression of the user as to quality of the service provided. The first parameter was measured by use of Transaction Log Files since they constitute a non-intrusive method enabling use to gather quantitative data in a reliable way. The second parameter was evaluated by means of short questionnaires which are used to gather the general impression on the product and its applications.

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:

- 95% of the users performed multilingual searches in the library database rather than monolingual ones;
- approximately 45% of the users felt whether satisfied or completely satisfied about the results of their search, while 9% of them were completely not satisfied;
- less than 25% of the unsatisfied users considered multilingual aspects not supported in TRANSLIB responsible for their dissatisfaction;
- approximately 85% of the users considered searching via the Multilingual Thesaurus more accurate and complete;
- over 55% of the users considered the translations of titles comprehensible;
- over 70% of the users considered the system easy to use from anyone, besides the experts;
- over 60% of the users considered TRANSLIB more friendly, useful and easy to understand than other OPACs (with no multilingual features) they had used in the past;
- 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 translations of titles. Hence, translations to and from English are more comprehensible than the ones 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 is consisted of titles in the local language and English. It has also to be underlined the remarkable satisfaction of the users from the use of Multilingual Thesaurus and from the overall search results they obtained by the system in contrast to other OPACs. Finally, the localization of the user interface was one of the main reasons the users favored this system.

5 Conclusions

Despite the fact that there are many OPACs supporting the automation of retrieving bibliographic information, only a few of them support real multilingual access to libraries catalogues. 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 stemmed from the integration of state-of-the-art already existing information tools and tools built from scratch. The integration of these tools was a hard task since the former were built for different purposes and, some of them, for different operation systems. For all these reasons practically libraries (e.g. public, private, university, etc.), book-trade enterprises, appropriate consultancy services as well as library schools will directly benefit by the presented system. From that aspect, TRANSLIB will 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.

Acknowledgments

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, Spanish Agency of International Cooperation, 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, unless so designated by other official documentation. The interested user may look for further information about TRANSLIB partners at the following URL address:
http://www.grial.uc3m.es/~aedo/translib.

References

Antworth E. *PC-KIMMO: A Two-Level Processor for Morphological Analysis*, Summer Institute of Linguistics, 1990.

ANSI/NISO Z39.50. *Information Retrieval (Z39.50)*, Application Service Definition and Protocol Specification, 1995.

Butcher R., *Multi-lingual OPAC Developments in the British Library*, Program, 27(2), pp. 165-171, 1993.

Buchinski E. J., W. L. Newman, and M. J. Dunn. *The Automated Authority Subsystem at the National Library of Canada*, J. Libr. Autom, 9(4), pp. 279-298, 1976.

Cousins S. A. and R. J. Hartley. *Towards Multilingual Online Public Access Catalogues*, Libri, 44(1), pp. 47-62, 1994.

Fluhr C., D. Schmit, P. Ortet, F. Elkateb, K. Gurtner, and V. Semenova. *Distributed multilingual information retrieval*, Proc. MULSAIC'96 Workshop, 1996.

Gibb F. *Knowledge-based Indexing in SIMPR: Integration of Natural Language Processing and Principles of Subject Analysis in an Automated Indexing System*, Journal of Document and Text Management, 1(2), pp. 113-125, 1993.

Gibb F. and G. Smart. *Knowledge-based Indexing: the View from SIMPR*, in C. MacDonald and J. Weckert (eds.), Libraries and Expert Systems, London, Taylor Graham, pp. 38-48, 1991.

Heid U. and J. McNaught. *Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerized Applications*, Final Report, 1991.

Hradilova J. *Thesaurus EUROVOC - Indexing language of the European Union*, Infocus, 1(3), pp. 66-69, 1995.

Hug H. and R. Noethinger. *ETHICS: An Online Public Access Catalogue at ETH-Bibliothek, Zurich*, Program, 11(2), pp. 133-142, 1988.

Kikui G., Y. Hayashi, and S. Suzaki. *Cross-lingual Information Retrieval on the WWW*, Proc. MULSAIC'96 Workshop, 1996.

Leeves J. (ed.) *Library Systems in Europe: a Directory & Guide*, Brussels-Luxemburg, 1994.

McAllistair C. *The Online Public Access Catalogue in DOBIS-LIBIS*, Program, 21(1), pp. 25-36, 1987.

Morris A. *The Application of Expert Systems in Libraries and Information Centres*, London, Bowker-Saur, pp. 37-62, 1992.

Pasanen-Tuomainen I. *Analysis of Subject Searching in the TENTTU Books Database*, Proceedings of the IATUL, 1, pp. 72-77, 1992(a).

Pasanen-Tuomainen I. *Monitoring Online Catalogues in the Nordic Technological University Libraries*, Nordinfo Nytt, 4, pp. 23-27, 1992(b).

Sgarbas K., N. Fakotakis and G. Kokkinakis. *A PC-KIMMO-based Morphological Description of Modern Greek*, Literary and Linguistic Computing, 10(3), Oxford University Press, New York, pp. 189-201, 1995.

Slater R. *Authority Control in a Bilingual OPAC: MultiLIS at Laurentian*, Library Resources and Technical Services, 35(4), pp. 422-458, 1991.

Synellis C. *TRANSLIB: User Survey*, Technical Report No 1.1, 1995.