UMIACS-TR-96-19 CS-TR-3615 April, 1996

A Survey of Multilingual Text Retrieval*

Douglas W. Oard
Electrical Engineering Department
and
Bonnie J. Dorr
Institute for Advanced Computer Studies and
Computer Science Department
University of Maryland, College Park, MD 20742
{oard|bonnie}@umiacs.umd.edu

Abstract

This report reviews the present state of the art in selection of texts in one language based on queries in another, a problem we refer to as "multilingual" text retrieval. Present applications of multilingual text retrieval systems are limited by the cost and complexity of developing and using the multilingual thesauri on which they are based and by the level of user training that is required to achieve satisfactory search effectiveness. A general model for multilingual text retrieval is used to review the development of the field and to describe modern production and experimental systems. The report concludes with some observations on the present state of the art and an extensive bibliography of the technical literature on multilingual text retrieval.

*The research reported herein was supported, in part, by Army Research Office contract DAAL03-91-C-0034 through Battelle Corporation, NSF NYI IRI-9357731, Alfred P. Sloan Research Fellow Award BR3336, a General Research Board Semester Award, and the Logos Corporation.

1 Introduction

In this report we survey the present state of the art in multilingual text retrieval. By "multilingual text retrieval" we mean the retrieval of documents (or, more precisely, electronic texts) based on explicit queries formulated by a human using natural language, regardless of the language in which the documents and the query are expressed. Neville has called this a "multilingually searchable system" [44]. Because the monolingual text retrieval problem has been well studied [28], we emphasize the cross-language aspect of text retrieval, the case in which queries are expressed in a language different from that of the documents.

Text retrieval is a process by which users seek to locate documents which contain information *about* the subject of their query.[1] The ubiquitous nature of electronic document preparation technology and recent improvements in page segmentation and text recognition technology have combined to make large amounts of text available in digital form, and dramatic reductions in digital storage and communications costs have made it practical to make those texts widely available. Text retrieval systems which are able to organize and retrieve material from this wealth of information are thus becoming increasingly important.

At present the vast majority of the demand for text retrieval is well satisfied by monolingual systems. One reason for this is that English is the *de facto* standard language of both commerce and science. Furthermore, when documents are desired in another language, it is often reasonable to expect the user to be able to formulate a query in that language. Nevertheless, there are important needs which cannot be satisfied by monolingual text retrieval systems. The examples which follow are meant to be illustrative, rather than exhaustive, but together they provide the principal motivation for this research.

- A collection contains documents in such a large number of languages that it would be impractical to form a query in each language.
- The documents themselves are expressed in more than one language. Consider, for example:
  - Technical documents in which English jargon appears intermixed with narrative text in another language.
  - Literary criticism which quotes substantial portions of a work in a different language.
  - Academic works which cite the titles of documents in different languages.
- The user is not sufficiently fluent in a document collection's language to express a query in that language, but is able to make use of the documents that are identified. This would certainly be useful for a user who is able to read but not to write well in the document collection's language, but there are a wide variety of circumstances in which a reader totally unfamiliar with the principal language of the document collection might find multilingual retrieval useful. For example:

[1] The text retrieval process is distinguished from the conventional database access paradigm by the user's desire to find documents *about* a subject rather than data which directly answers the query. A conventional database of bibliographic records can be used to perform text retrieval, but other approaches are also possible.

– A collection of images that are indexed by captions in a language that is unfamiliar to the user.
– A researcher seeking to determine which individuals and institutions have conducted research on a particular topic.
– A user with sufficient resources to translate the selected documents into a language that he or she is able to understand.

This last example points up a synergistic relationship between machine assisted translation and multilingual text retrieval. Multilingual text retrieval can be used to the number of documents requiring translation, while machine assisted translation makes it practical to translate the selected documents at a reasonable cost. Incremental improvement in either technology should result in a greater demand for both. A similar relationship exists between multilingual text retrieval and fully automatic machine translation. Although (except in narrow domains such as weather reporting) translations produced by fully automatic systems are of significantly lower quality than machine assisted translations, they can be used in a "screening" role during document selection [43].

Figure 1 illustrates how fully automatic and machine assisted translation resources could be integrated with a multilingual text retrieval system. With such a system, queries can be constructed in whatever language the user finds convenient, and documents will be returned in whatever language they are expressed. If necessary, fully automatic machine translation can be used to produce screening-quality translations that allow the user to select documents. When a higher quality translation is required, selected documents can be submitted for machine assisted human translation.

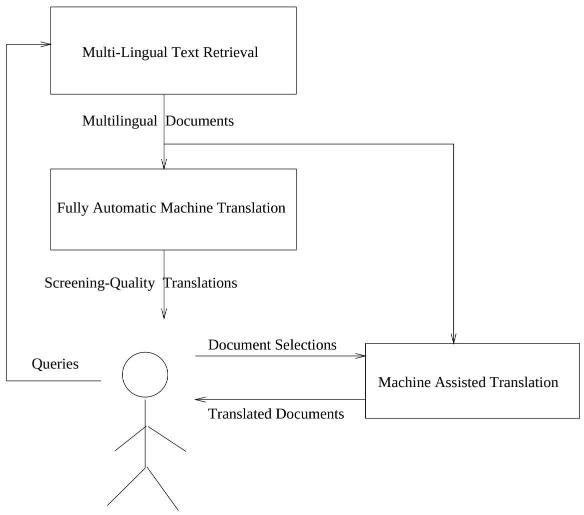

Figure 1: Integrating multilingual text retrieval with machine translation.

Before proceeding it might be useful to identify related research that is outside the scope of this survey. The term "multilingual" is also commonly used to refer to text retrieval systems which can be parameterized to search in *one* of several languages (c.f. [14]). In such systems both the query and the documents must in the same language, so such systems are actually monolingual text retrieval systems. It is possible to use several monolingual text retrieval systems to retrieve documents from a multilingual document collection, but we do not consider such an approach multilingual text retrieval in the sense of our original definition.

Occasionally, "multilingual" is used even more broadly to describe features of the user interface that allow text to be entered and/or displayed using more than one language or character set(c.f. [56]). This concept is also referred to as "localization" or "internationalization" of software, reflecting the motivation behind the design of a linguistically parameterized user interface. In this context, an online library catalog might be described as "multilingual" if it allowed the user to select the language in which help screens are displayed, even if only monolingual searching is possible.

These closely related research areas offer important perspectives on text retrieval in languages other than English that would be useful to developers of truly multilingual text retrieval systems. Many components of a multilingual text retrieval system, such as character coding, font construction, morphology, and phrase recognition, can be initially investigated in the context of monolingual text retrieval and then later applied to multilingual text retrieval. But our interest is in *cross-language* text retrieval. So in this survey we restrict our attention to techniques for selecting documents in one language based on queries expressed in another, and we subsequently use the term "multilingual text retrieval" to mean exactly that.

2 Text Retrieval System Model

The goal of a text retrieval system is to present the user with a set of items that will satisfy his or her information need. We refer to the concrete expression of the information need in words as a "query," and we call the items from which we select "documents." Every approach to text retrieval has two basic components: some technique for representing texts (i.e., queries and documents), and some way of comparing those representations. The objective is to automate the process of examining documents by computing comparisons between the representation of a query and the representations of the documents. This automated process (referred to as "text retrieval") is successful when it produces results similar to those produced by human comparison of the query with the documents.

This basic text retrieval model is often extended to account for observed differences in the characteristics of queries and documents. For example, queries are often quite short (with lengths of one or two words not being uncommon), while documents might easily be hundreds of pages long. Another issue is that users frequently adopt a vocabulary that differs significantly from that in the documents that contain the information they seek [29]. This is known as the "paraphrase problem." One way that text retrieval systems accommodate such differences is by constructing representation functions that treat queries and documents differently to arrive at compatible representations. This distinction also provides the basis for multilingual text retrieval, which is simply a special case of the paraphrase problem [25], so we spend a moment to formalize this idea.

Figure 2 depicts the representation and comparison process graphically. Formally, the domain of the query representation function q is Q, the collection of possible queries and it's range is R, the unified space of text (i.e., query and document) representations. The domain of the document representation function d is D, the collection of documents, and its range is R.[2] The domain of the comparison function c is $R \times R$ and its range is [0,1], the set of real numbers between zero and one. In an ideal text retrieval system,

$$c(q(\text{query}), d(\text{doc})) = j(\text{query}, \text{doc}), \forall \text{query} \in Q, \forall \text{doc} \in D,$$

where $j : Q \times D \mapsto [0, 1]$ represents the user's judgement of some relationship between two texts, measured on a single ordinal scale (e.g., content similarity or style similarity). Figure 2 illustrates this relationship.

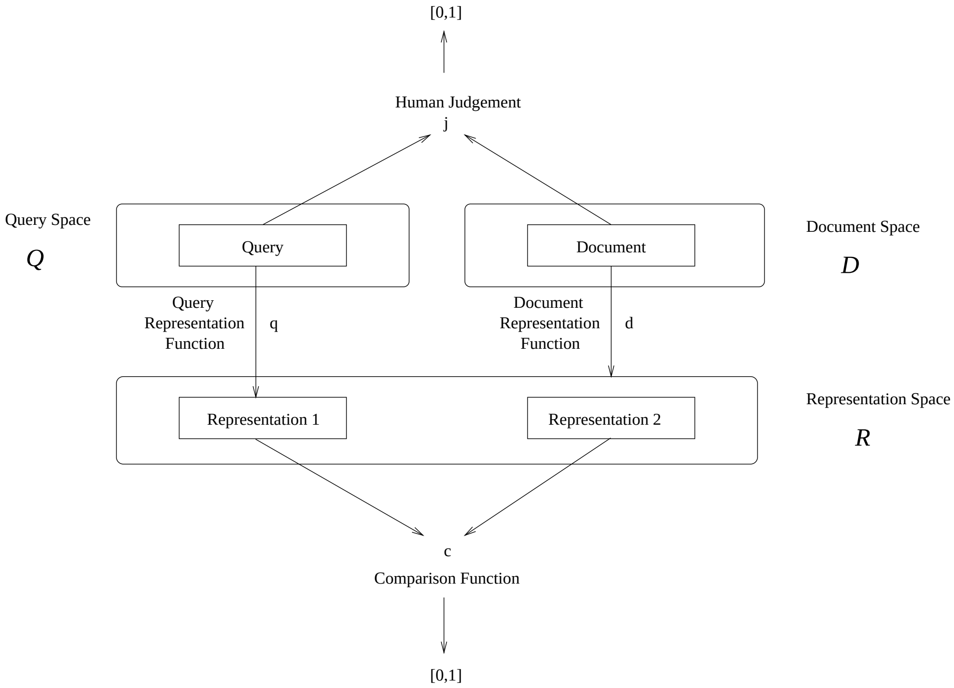

Figure 2: Text retrieval system model.

In this survey we describe two types of text retrieval systems: exact match and ranked retrieval. The text retrieval model presented above can be specialized to describe either approach. In an exact match text retrieval system the range of c is restricted to be either zero or one, and it is interpreted as a binary judgement about whether a document satisfies the Boolean expression specified by the query. Exact match text retrieval systems typically provide an unranked set of documents which satisfy the user's query, and most existing multilingual text retrieval systems fall into this category. We describe their operation in some detail in section 3.2.

In ranked retrieval the system attempts to impose a total order on the documents in such a way that the most useful documents are near the top of the list.[3] Three types of ranked retrieval systems are described in this survey. In all three the range of c is [0,1],

[2] The document representation function's effect is often referred to as an "indexing" because the results of applying d to each document in the collection are often used to construct an index of some sort to improve query-time efficiency.

[3] Ranked retrieval systems which construct only a partial order have been proposed, but we are aware of no work on multilingual text retrieval research in which such a model is used.

Selected as	Actually is: Relevant	Actually is: Not Relevant
Relevant	Found	False Alarm
Not Relevant	Missed	

$$\text{Precision} = \frac{\text{Found}}{\text{Found} + \text{False Alarm}}$$

$$\text{Recall} = \frac{\text{Found}}{\text{Found} + \text{Miss}}$$

Table 1: Measures of text retrieval effectiveness.

but they differ in how this "retrieval status value" is interpreted. In a "ranked Boolean" retrieval system the value is interpreted as the degree to which one text satisfies the Boolean expression specified by the other. "Probabilistic" retrieval systems generalize this concept still further, interpreting the value as the probability that a text is relevant to a query. Many probabilistic retrieval systems have been designed to accept queries expressed in natural language rather than as a boolean expression. In "similarity-based" retrieval systems (such as the "vector space" approach), the retrieval status value is interpreted as measuring the degree to which the content (or some other aspect) of two texts is similar.[4]

Real text retrieval systems at best approximate this ideal, and the field of text retrieval system evaluation is devoted to characterizing how close that approximation is.[5] A common simplifying assumption (and one which is quite natural for exact match retrieval systems) is that j is binary valued and is given. In other words, it is assumed that documents are either relevant to the query or they are not, and that this "relevance judgement" can be reliably ascertained by a user. Under this assumption the effectiveness of an exact match retrieval system is typically characterized by two statistics, "precision" and "recall."[6] Precision is the fraction of the selected documents which are actually relevant to the user's information need, while recall is the fraction of the actual set of relevant documents that are correctly classified as relevant by the text retrieval system.[7] Viewed another way, precision is one minus the false alarm rate and thus measures accuracy, while recall measures how complete the search is. Table 1 illustrates these relationships.

Evaluation of ranked retrieval systems is more complex. One common effectiveness

[4]The fundamental assumption of similarity-based text retrieval is that documents with content similar to the content of the user's query will satisfy his or her information need.

[5]The quality of the approximation is a measure of effectiveness. Efficiency and usability are other important aspects of text retrieval system evaluation.

[6]Other effectiveness measures have been proposed (c.f. [69]), but precision and recall are the most commonly reported statistics.

[7]Relevant documents are those which could contribute to fulfilling the information need. For a comparison of relevance with "utility" see, for example, [69].

measure for such systems is "average precision" [65]. It is computed by choosing successively larger sets of documents from the top of the ranked list that result in evenly spaced values of recall between zero and one. Five (0.1, 0.3, 0.5, 0.7, 0.9), nine or eleven recall points are typically used. Precision is then computed for each set. The process is repeated for several queries, and the mean precision for each reported value of recall computed. The mean of these numbers is then computed and reported as a single figure of merit for the system. Larger values of average precision are better, and comparisons are only meaningful when the same collection of queries and documents are used. Average precision does, however, obscure variations across queries with different characteristics (such as having differing numbers of relevant documents in the collection). Furthermore, because the density of relevant documents is (hopefully) highest near the top of the ranked list, precision typically declines each time the set of documents is expanded to increase recall.[8] Full precision-recall plots can be useful when the decay rate of precision with increasing recall differs so markedly that each system outperforms the other for some value of recall.

3 Approaches to Multilingual Text Retrieval

Building on recent work by Fluhr [25], we next present a taxonomy of multilingual text retrieval approaches. Three main themes have emerged in the research literature: thesaurus-based approaches, corpus-based approaches and modular use of machine translation for text translation. We begin by describing the text translation approach both because it is straightforward and because its weaknesses help to motivate investigation of techniques which more closely integrate the translation and retrieval functions.

3.1 Text Translation

Perhaps the most straightforward approach to multilingual text retrieval is to implement either q or d using a fully automatic machine translation system in order to bring the query and the document into a representation space R that is based on a single language. Surprisingly, although this approach has been suggested repeatedly in recent years [16, 25, 47, 72] we are aware of only one experiment for which results have been reported [26, 25, 57, 58].

One weakness of present fully automatic machine translation systems is that they are able to produce high quality translations only in limited domains. Fluhr observes that text retrieval systems are typically more tolerant of syntactic than semantic translation errors, but that semantic accuracy suffers when insufficient domain knowledge is encoded into a translation system [25].[9] Since encoding domain knowledge can be expensive, Fluhr's observation would suggest that the effectiveness of a machine translation approach to multilingual text retrieval will be limited, particularly when it is the relatively short queries that are translated.

[8] Because there may be fewer relevant documents than recall points, step function interpolation is used to calculate the precision for the chosen values of recall. By convention, the step function chosen for interpolation decreases monotonically, thus *ensuring* that precision never increases with increasing recall.

[9] The component of a machine translation system which encodes domain knowledge is typically referred to as a lexicon.

It might be possible to partially mitigate this problem by translating the documents rather than the queries. Because the documents are typically much longer than queries, a machine translation system embedded in d would have considerably more contextual information on which to base semantic choices than one embedded in q. Furthermore, text retrieval systems are typically tolerant of occasional semantic inaccuracy if the dominant pattern of the semantic choices is appropriate. Longer documents usually include a larger vocabulary, and a large vocabulary could improve the prospects for developing a dominant pattern of correct semantic choices.

However, the efficiency of available machine translation becomes an issue when a translation system is embedded in d, because d typically must be applied to a very large number of documents. Moreover, some of the work done by a machine translation system yields no improvement in retrieval effectiveness. For example, translation of text requires choosing word order and adding closed class words in the target language.[10] But both of these features are typically removed by q and d.

In fact, some of the work done by a machine translation system could actually reduce some measures of retrieval effectiveness. Because word senses may not be grouped with words in the same way in different languages, machine translation systems attempt to make the best possible determination of the sense in which polysemous words are used.[11] Following that analysis a single sense is chosen for each polysemous word. In a text retrieval system, however, q and d can be designed to preserve information about uncertainty and c can be designed to exploit that information to improve effectiveness. As a simple example of this, an exact match text retrieval system could substitute *every* possible translation for a polysemous word, thus increasing recall (at the expense of precision). Some types of ranked retrieval systems are able to represent and exploit information about the probability that each sense of a polysemous word is correct. If this information could be extracted from the machine translation system, average precision might be improved by increasing recall while limiting the adverse effect on precision.

These observations suggest that when designing q and d functions for multilingual text retrieval, the type and depth of processing should be determined by the ability of the representation space R to represent the results of that processing and the ability of the comparison function c to use that information. We could either constrain our processing by the ability of existing techniques to use the resulting information or we could design new representations and comparison functions to exploit the information that machine translation technology can provide. In the remainder of this section we will describe how these two approaches have been integrated in both practical and experimental systems.

3.2 Multilingual Thesauri

In this survey we define a thesaurus as a tool which organizes terminology to encode domain knowledge for use by an application. Thus a thesaurus is an ontology that is specialized to organizing terminology. A multilingual thesaurus is one which organizes terminology from more than one language. Bilingual dictionaries, which typically define

[10] Closed class words, words which carry little content, are typically removed by the "stopword" list in a text retrieval system.

[11] Polysemous words are words which have more than one meaning.

Thesaurus Type	Characteristics
Subject Thesaurus	Hierarchical and associative relations. Unique term assigned to each node.
Concept List	Term space partitioned into concept classes.
Term List	List of cross-language synonyms.
Lexicon	Machine readable syntax and/or semantics.

Table 2: Examples of multilingual thesauri.

terms with respect to other terms, are clearly subsumed by this definition.[12] Lexicons in computational linguistics, which encode syntactic and semantic information about terms, are included as well. Complex thesauri used as a concept index in automatic text retrieval systems, are also within the scope of our definition of a thesaurus. Even a simple bilingual listing of technical terms in which each term is assigned a unique translation, would be a thesaurus by our definition. We realize that this is an unusually broad definition of the term "thesaurus." But because no standard terminology succinctly captures the concept we describe, we have chosen to use the term most closely associated with present multilingual text retrieval practice. Table 2 shows some common types of thesauri used in multilingual text retrieval systems.

Thesaurus-based techniques share certain advantages and limitations. Because thesauri can represent relationships between terms and concepts in a way that humans find understandable, thesaurus-based text retrieval allows users to exploit insight gained during the search process to reformulate better queries. Furthermore, because a significant amount of domain knowledge can be encoded in the thesaurus, in the hands of a skilled user a thesaurus-based text retrieval system can be a powerful tool. On the other hand, use of a thesaurus imposes an *a priori* limitation on both the vocabulary the user may employ and on the domain to which the text retrieval system can be applied.[13] Present techniques for thesaurus construction and maintenance are resource-intensive, and the training and effort required to effectively use the concept relationships contained in a sophisticated thesaurus can be substantial. We discuss some of these limitations in more detail at the end of section 3.2.2 after we have described how thesauri are used for multilingual text retrieval.

Several aspects of domain knowledge can be encoded in a thesaurus. The key feature of every multilingual thesaurus is a specification of cross-linguistic synonymy.[14] Hierarchical concept relationships (broader term, narrower term) and associative relationships (related term, synonymous term) are typically included in more sophisticated thesauri.[15]

Thesauri can be used either manually or automatically. In so-called "controlled

[12]As used in this survey, "terms" is used inclusively to mean either single words or multi-word phrases.

[13]Even fairly comprehensive dictionaries lack detailed coverage of a large number of domains, an observation confirmed by the development of countless specialized technical dictionaries.

[14]The specification of cross-linguistic synonymy need not be complete because some terms may not have direct translations in another language.

[15]Systems which do not make the thesaurus accessible to the user may use only an internal representation for nodes in a conceptual hierarchy, so the "broader terms" we refer to may not be intended for human use.

vocabulary" systems, every concept is labeled with a unique descriptive term so that the user can manually specify the appropriate concepts in his or her query. When the concept relationships encoded in a thesaurus are used automatically, the technique is often referred to as "concept retrieval." In a simple concept retrieval system a concept list could be used to replace each term with its concept class to increase recall (again at the expense of precision). A more sophisticated approach, known as "query expansion" would be to use the concept relationships encoded in the thesaurus to choose terms that could improve both precision and recall. We give examples of both techniques below.

Both concept substitution and query expansion represent attempts to increase recall by reducing the effects of the paraphrase problem. Precision can be increased by including syntactic or semantic information in a thesaurus to mitigate the effects of polysemy.[16] For example, in a controlled vocabulary system semantic information (called a "scope note") is often provided in the thesaurus to help users manually choose the correct term. A concept retrieval system could apply this idea by automatically tagging some words with their part-of-speech and then select translations that are appropriate for that part-of-speech. We describe such a system below.

We begin our discussion of thesaurus-based systems with a description of two important early experiments that demonstrated the potential of that approach. We will then describe developments in controlled vocabulary and concept retrieval systems, followed by a description of projects which have exploited encoded semantic knowledge.

3.2.1 Early Work

Pigur describes a multilingual controlled vocabulary thesaurus in English, French and German that was developed for the International Road Research Documentation (IRRD) system in 1964 [53]. But the earliest reported experimental results on the effectiveness of multilingual text retrieval were reported by Salton at Cornell University in 1969 [64]. Salton augmented his SMART text retrieval system[17] with a multilingual concept list constructed by translating some of the words in an existing English concept list into German. Forty-eight English queries for a collection of library science abstracts were manually translated into German, and all four possible language pairs were evaluated. On the 468 German abstracts, the use of English rather than German queries reduced the average precision[18] from 0.35 to 0.34 (3%),[19] while on 1095 English abstracts the use of German rather than English queries reduced the average precision from 0.33 to 0.31 (6%). From this Salton concluded that although retrieval effectiveness varied across document collections (a well known phenomenon in text retrieval), "cross-language processing ...is nearly as effective as processing within a single language." After examining the retrieval failures in more detail he concluded that "it would therefore seem essential that a more complete thesaurus be used under operational conditions for future experiments."

[16] Polysemy resolution is often referred to as "word sense disambiguation."

[17] SMART is a vector space ranked text retrieval system.

[18] In these studies Salton reported precision at five values of recall evenly spaced between 0.1 and 0.9.

[19] We report average precision to two decimal places, but do not mean to imply that the results are statistically significant to two figures. We report the percentage difference based on these values with reference to the monolingual technique in an attempt to facilitate comparison with other approaches.

For a 1973 paper Salton implemented an English-French multilingual concept list, this time achieving more complete coverage by independently developing the section for each language after establishing a common set of concepts [63]. Again, no information about the relationships between concepts was encoded or used. In this study he obtained a French-English parallel corpus of 52 abstracts about documentation and used a set of 16 translated queries.[20] Salton observed that on French abstracts the use of English rather than French queries *increased* the average precision from 0.43 to 0.45 (5%) but that on English documents the use of French rather than English queries decreased the average precision from 0.43 to 0.38 (12%). This last result is perhaps explained by the sensitivity of the average precision metric to the rank assigned to a single abstract in such a small collection (a speculation reinforced by the nearly step-function shape of the precision-recall graphs in this case). Salton observes, however, that the smaller English vocabulary in this domain also gave English queries the advantage of operating at a somewhat higher level of abstraction.

At about the same time, Pevzner performed a similar experiment using the Russian PNP-2 exact match controlled vocabulary text retrieval system [52].[21] Pevzner expanded the PNP-2's sophisticated Russian thesaurus, which contained several thousand words, several thousand concepts, and over 600 relationships between those concepts, to English [51]. He then used PNP-2 to retrieve both Russian and English documents based on an identical set of 103 short Russian queries.[22] Using quantities he calls "losses" and "noise," Pevzner reported that a sign test revealed no statistically significant difference (to 95% confidence) between selections from 4000 Russian and 4400 English electrical engineering documents.[23]

3.2.2 Controlled Vocabulary Systems

By 1973 it was well established that both controlled vocabulary and concept retrieval systems with multilingual thesauri could achieve performance across languages on a par with the within-language performance of the same techniques. Commercial acceptance soon followed, and by 1977 Iljon was able to identify four multilingual text retrieval systems operating in Europe [34]. Since this early work, six principal lines of research on multilingual thesauri have emerged: design standards, development and maintenance tools, special purpose hardware, new language pairs and domains, user interfaces, and user needs assessment.

In 1970 it was already becoming clear that standardization of thesaurus development to prevent "creation of many divergent and incongruent subject indexing vocabularies" would be beneficial, and in 1971 the United Nations Educational Scientific and Cultural Organization (UNESCO) proposed standards for multilingual thesaurus development [73]. In 1973 the International Standards Organization (ISO)[24] took up the matter, and by 1976 the draft specification had been greatly expanded [3]. Approved in 1978 as ISO 5964 and most recently modified in 1985, the standard describes how

[20] A parallel corpus is a collection of documents in which every document is translated into every language.

[21] PNP-2 stands for "Pusto-Nepusto-2." In his 1973 paper [63], Salton translates the name of Pevzner's system as "Empty-Nonempty 2" and transliterates Pevzner's name as "Pevsner."

[22] The examples Pevzner provides are all between 2 and 5 words.

[23] Unfortunately, the cited definitions of "losses" and "noise" are in Russian, and Pevzner's summary of their definition appears to be incomplete.

[24] ISO Technical Committee 46, Working Group 5.

domain knowledge can be incorporated in multilingual thesauri and identifies alternative techniques for multilingual thesaurus development. In 1982 the Soviet Union adopted a similar standard, GOST 7.24-80 [49].[25]

The European Parliament's EUROVOC is an example of a modern ISO 5964 multilingual thesaurus [27]. First published in 1984, EUROVOC now includes all nine official languages of the European Community, and portions of it have been translated into additional languages (c.f. [13]).[26] Thesaurus design remains expensive, and this fact has limited the domains to which which controlled vocabulary retrieval has been applied. But EUROVOC demonstrates that once the basic concept relationships have been defined for a domain, extension of an ISO 5964 multilingual thesaurus to additional languages is quite practical.

As large multilingual thesauri have proliferated, design and maintenance tools have become increasingly important. In 1970, Neville described a procedure for merging thesauri that could be used to merge monolingual thesauri to produce a multilingual thesaurus [45], and in 1975 he contrasted this approach with other ways of producing multilingual thesauri [46]. Bollmann and Konrad presented a technique for merging monolingual with bilingual thesauri in 1975 [7], and in 1977 Iljon surveyed available thesaurus design and maintenance tools and described the operation of the Commission for the European Communities' ASTUTE system [33].[27]

More recently, an automatic technique for using a thesaurus to generate corresponding indexing terms in four languages was described by Pelissier and others in 1986 [50]. In 1987 Kalachkina presented an algorithm for merging thesauri in different languages [35] and in 1989 Loginov described tools developed in the Soviet Union to maintain a Russian-English version of the (monolingual) United States National Library of Medicine's Medical Subject Heading (MeSH) thesaurus [41]. Loginov's paper illustrates a case in which external factors (changes to MeSH) generate the thesaurus maintenance requirements. Sogoaga of SABINI, a Spanish library automation company, also described the design of interactive tools for multilingual thesaurus maintenance [19]. The SABINI system was designed for automation of bibliographic records in an online library catalog. Sogoaga provided no examples of implementations for specific languages, however.

In 1988 Kitano, from NEC's Tokyo Software Engineering Development Laboratory, described the development of a hardware tool designed to support multilingual text retrieval [36]. He implemented a Japanese-English thesaurus using a NEC integrated circuit known as the "Intelligent String Search Processor." At the time, the ISSP thesaurus implementation had not been integrated with a text retrieval system, however, so no experimental results were reported.

The research literature on multilingual text retrieval offers several examples of systems which have implemented new language pairs [2, 10]. and new domains [4, 39, 75]. Because this type of report can describe the effect of previously unseen linguistic phenomena on thesaurus design and other aspects of a text retrieval system (e.g. stemming and compound recognition), case studies can provide useful insights into the complexity

[25]BS 6723, DIN 1463 and AFNOR NF Z 47-101 are the national standards for multilingual thesaurus development in the United Kingdom, Germany and France, respectively.

[26]The nine languages are Danish, Dutch, English, French, German, Greek, Italian, Portuguese, and Spanish.

[27]ASTUTE stands for Automated System for Thesaurus Updating, Testing and Editing.

of implementing ISO 5964 and similar national standards.

Semturs, of IBM Netherlands' Scientific and Cross Industry Center, provided some insight into the contemporary commercial development of user interfaces for multilingual text retrieval systems in the mid-1970's [66, 67]. He described the capabilities of a commercial product, the STAIRS-TLS exact match text retrieval system, which was able to accommodate queries and documents in German, English and French.[28] STAIRS was originally a monolingual full-text retrieval system,[29] and STAIRS-TLS added a multilingual thesaurus. It included an interactive interface with thesaurus-based tools to facilitate controlled vocabulary query formulation. Semturs' papers report no performance figures, but they offer some insight into the market demands for multilingual text retrieval.

More recently, a team at the University of Huddersfield Centre for Database Access Research in the United Kingdom led by Pollitt has integrated multilingual thesauri with interactive personal computer technology to address one of the fundamental limitations of controlled vocabulary text retrieval [6, 40, 54, 55]. Experience has shown that although the domain knowledge that can be encoded in a thesaurus permits experienced users to form more precise queries, casual and intermittent users have difficulty exploiting the expressive power of a traditional query interface in exact match retrieval systems. Adapting their Menu-based User Search Engine (MenUSE) to use the European Parliament's multilingual EUROVOC thesaurus, Pollitt's team has developed a query formulation tool which facilitates visual browsing in the user's preferred language. Pollitt's team has also extended the English thesaurus for the INSPEC database to Japanese and integrated it with MenUSE. The cited works do not report experimental results on the utility of the multilingual MenUSE interface, but a monolingual evaluation of MenUSE on the INSPEC database is presented in [68].

Controlled vocabulary text retrieval systems are widely used in libraries, and user needs assessment has received considerable attention from library and information science researchers. Rolling described a user needs assessment conducted for the Council of the European Community in 1974 [62], and the TRANSLIB project, a part of the European Commission's I*M-Europe Telematics for Libraries program, provides a recent example of user needs assessment [71]. TRANSLIB's goal is development of a trilingual (Greek, Spanish and English) subject search capability for an online library catalog. Chachra discussed user needs assessment for multilingual online library catalogs in [12] more generally, and provided examples from the VTLS online library catalog system. In addition to monolingual full text searching, VTLS used a multilingual thesaurus to suggest controlled vocabulary search terms in a second language. Rolland-Thomas described a similar feature in the Canadian DOBIS bilingual online library catalog, and discussed the utility of more automatic techniques from a user needs perspective [61].

Pasanen-Tuomainen, of the Helsinki University of Technology, reported results from a usability assessment for a multilingual online library catalog, TENTTU, that incorporated both multilingual controlled vocabulary and monolingual full text searching. [48][30] Examining 2,620 search commands issued during 655 sessions, she found that

[28]STAIRS-TLS stands for STorage And Information Retrieval System-Thesaurus and Linguistic integrated System.

[29]A full-text retrieval system is one which can index any word appearing in any document, regardless of whether it appears in a thesaurus.

[30]TENTTU used the Universal Decimal Classification (UDC), a greatly expanded version of the Dewey

library staff used the controlled vocabulary in 46 of their 337 search commands (14%), but that other patrons used it for less than 3% of their commands. She did observe, however, that 11% of the remaining search commands contained words found in the thesaurus that could have been mapped across languages had TENTTU been designed to do so. She also suggested that limited thesaurus availability and inadequate patron training might have reduced thesaurus utilization.

Multilingual text retrieval systems are widely used today, but nearly every commercial system that we are aware of uses an exact match approach.[31] Sophisticated multilingual thesauri have been developed for many domains and many languages, and the procedures for adding new domains and languages are well understood. Before describing experimental approaches, then, we pause to outline what we hope to gain by examining other techniques. After all, if we are to improve on present practice we must understand the limitations of present systems. Three key factors bear examination in this regard: cost, usability by untrained users, and effectiveness [22].

Thesaurus construction is an expensive activity. But thesaurus use can be even more expensive because in a controlled vocabulary system every document must be assigned terms that reflect the concepts it contains.[32]. Although automated tools can improve human productivity, as long as human intellectual activity is required to recognize and organize information the costs will remain substantial. In fact, with the sustained dramatic decline of computer hardware costs, human activities such as thesaurus maintenance and controlled vocabulary indexing have come to dominate system costs. This limits both the scalability of existing thesaurus-based systems to accommodate the rapid growth in electronically accessible texts and the generalizability of the technique to new domains (e.g. personal document collections) for which construction and/or use of a thesaurus is economically impractical.

Another important limitation of controlled vocabulary text retrieval techniques, and one which is shared by full text exact match techniques as well, is that untrained users seem to have difficulty exploiting their capabilities. Significant differences between the performance of skilled and untrained users have been observed with their choice of terms, their use of the term relationships that can be encoded in a thesaurus, and their use of operators such as *and*, *or* and *not* for query construction. In many cases it has proven more economical to provide trained intermediaries than to provide adequate training to each user. Advanced user interfaces such as MenUSE offer some potential for mitigating this problem, and expert systems that construct Boolean queries from natural language have been investigated in a monolingual context [42]. The ranked retrieval techniques we describe in section 3.3 represent another approach to solving this problem. Ranked retrieval systems typically accept queries in natural language and allow a (relatively) unconstrained choice of terms. In general, the goal of ranked retrieval is not to replace exact match techniques but rather to augment them with techniques that improve the search effectiveness of untrained users. In multilingual text retrieval, ranked retrieval techniques also allow us to avoid an unsolved problem identified by Chachra [12], who observed that single terms in one language can correspond to complex boolean expressions in another when a controlled vocabulary is not

decimal system, as a multilingual subject thesaurus.

[31]The exception is the SPIRIT system developed for EMIR which we discuss below.

[32]Dubois discounted this factor, but his analysis was conducted in the context of abstracting services in which the cost of abstract preparation dominates the processing cost for newly arrived documents

used.

A third reason to investigate corpus-based techniques is to improve effectiveness. Language use is a creative activity, and new words enter human languages each year. Because thesaurus construction is time-consuming, thesauri in production applications necessarily lag somewhat behind the common use of terminology. Furthermore, there is some evidence that thesaurus designers have more difficulty anticipating which concepts and relationships will be useful to their system's eventual users than a cursory inspection of the thesaurus would suggest [65].[33] Since corpus-based techniques are based on the observed statistics of term usage, they offer some hope that important aspects of current term usage can be identified and exploited. The potential of corpus-based multilingual text retrieval techniques has yet to be realized in a large-scale experiment, however, so we will begin our discussion of experimental techniques with those which include some form of human-usable thesaurus.

3.2.3 Concept Retrieval

Salton's early experiments provide one example of concept retrieval. An alternative to Salton's representation of concepts in R is to represent terms, using the multilingual thesaurus to guide the term selection process. This is a variation on query expansion, a well studied technique for monolingual text retrieval [9].[34] The basic idea of query expansion is to accommodate term usage variations by augmenting the terms in the query with related terms. But because query expansion typically improves recall at the expense of precision, selection of inappropriate terms could reduce overall performance measures such as average precision. So, in the context of multilingual text retrieval, the goal of query expansion techniques is to accommodate *cross-linguistic* term usage variation while minimizing the adverse impacts on effectiveness.

Recently, Davis and Dunning of New Mexico State University have evaluated several multilingual text retrieval techniques, one of which is based on query expansion [16]. For the evaluation of Spanish text retrieval at the fourth Text Retrieval Conference (TREC-4) they manually translated 25 Spanish queries into English and then used them to select documents from a collection of 58,000 Spanish articles from the Mexican newspaper "El Norte" using the INQUIRY text retrieval system. They then automatically formed a Spanish query by selecting every English translation for each word in the query from a simple bilingual term list.[35] This approach, which they used as a benchmark against which to compare their corpus-based approaches, achieved an average precision of 0.04.[36] Five of the ten participants in the TREC-4 Spanish text retrieval evaluation achieved an average precision exceeding 0.21 on the same collec-

[33] Discussions about the relative effectiveness of controlled vocabulary and statistical text retrieval are often marked by considerable enthusiasm on both sides, however, so it is difficult to find impartial evaluations on this issue.

[34] The unique feature of cross-language query expansion is that the original term is removed from the expanded query unless it carries the same meaning in both languages.

[35] Davis and Dunning used an online version of the Collins English-Spanish dictionary as a bilingual term list.

[36] Average precision in TREC is computed over eleven points evenly spaced between 0.0 and 1.0. Documents selected by any participating system are evaluated and the remaining documents are not examined. "Unknown" documents are treated as "not relevant," so TREC actually lower bounds recall. That bound is thought to be fairly tight.

tion by using the Spanish queries directly, so Davis and Dunning's results suggest that unconstrained query expansion is of limited value for multilingual text retrieval.[37]

Building on this work, Hull and Grefenstette at Rank-Xerox in France have evaluated the potential of more sophisticated approaches to query expansion [32]. They manually translated 50 short TREC queries[38] into French and created a bilingual term list that contained every possible translation for each French word.[39] Unconstrained cross-language query expansion was then used to select from approximately 500,000 newspaper articles for which relevance judgements were available using the SMART vector space text retrieval system.[40] They found that adding phrases[41] to the bilingual term list increased their effectiveness measure[42] from 0.27 to 0.36 (33%).[43] Using the original English queries, Hull and Grefenstette achieved an effectiveness measure of 0.39. From this they concluded that inclusion of phrases in a bilingual term list can allow the query expansion technique to perform almost as well across languages as traditional statistical techniques do in a monolingual setting.[44]

The European Multilingual Information Retrieval (EMIR) project, led by Fluhr of the French Institut National des Sciences et Techniques Nucléaires (INSTN), also used a query expansion technique [26, 57, 58, 59, 70]. An ESPRIT II[45] project, EMIR work proceeded between November of 1990 through March of 1994. The goal of EMIR was to extend the SPIRIT text retrieval system (which was originally developed by Fluhr and others) to multiple languages.[46] The initial language pair was English and French, and it was later extended to German. Analit Ltd., a Russian company, is extending SPIRIT to Russian. SPIRIT is a ranked Boolean text retrieval system, in which sets are selected using successively smaller portions of the original query and then ranked for display in order of increasing generality.

For the French/English language pair there were 33,153 mappings from French terms to one or more English terms. Each such mapping had between 1 and 24 possible English terms, and the median number of English terms for a French term was 2. English terms which did not appear in the document collection were then eliminated. On a parallel bilingual corpus from the European Court of Justice, this achieved at least a 40% reduction in the number of target terms for 92.6% of the transfer rules. More comprehensive performance results are given below.

---

[37] One of the systems exceeding an 0.21 was submitted for comparison by Davis and Dunning. The best average precision achieved by a monolingual system was 0.49.

[38] Hull and Grefenstette used shortened versions of TREC queries 51-100 which had an average length of seven words.

[39] The bilingual term list was manually constructed using the third edition of the Robert and Collins French-English dictionary.

[40] Use of existing TREC relevance judgements lower bounds both precision and recall.

[41] Only phrases appearing in the same dictionary were added.

[42] Hull and Grefenstette reported precision averaged over fixed size sets containing the top ranked 5, 10, 15, and 20 documents.

[43] These figures were used by Hull and Grefenstette as a benchmark for evaluating automatic techniques for constructing term lists from an online dictionary that was designed originally for human use.

[44] In comparing these results with those of Davis and Dunning it is important to consider that Hull and Grefenstette selected their effectiveness measure with interactive applications in mind.

[45] ESPRIT II was the second phase of the European Commission's information technology research program.

[46] SPIRIT stands for Syntactic and Probabilistic Indexing and Retrieval of Information in Text.

3.2.4 Encoding Semantic Information

Another aspect of the EMIR project was application of fast but shallow parsing to exploit semantic information that was encoded in the thesaurus [59, 57, 58, 70]. The number of English terms was reduced by labeling each English term with the corresponding part of speech and then only choosing those English terms which were appropriate for the syntactic usage of the French term. The EMIR thesaurus was a bilingual term list in which semantic information, encoded as compounds, was used in place of concept relationships. In EMIR, terms included words, phrases and compounds. Because compounds link key terms together on the basis of their semantic relationship rather than their surface form, compound formulation is more powerful than simple phrase extraction. Because the order of the components in a compound was sometimes switched in the target language, the term list entries for compounds were constructed to account for transposition when necessary.

The EMIR version of SPIRIT was evaluated on the Cranfield collection of 1398 aeronautical abstracts using 225 queries which had been translated into French by the French Army Documentation Center. English documents were retrieved in response to French queries. For comparison, the French queries were translated back into English using the SYSTRAN fully automatic machine translation system and documents were selected using a monolingual version of the SPIRIT text retrieval system. EMIR increased average precision over the combination of SYSTRAN and SPIRIT from 0.21 to 0.27 (29%), but use of the original English queries with SPIRIT further increased average precision to 0.34 (26%).[47]

Some more exploratory projects with potential multilingual text retrieval applications have also been reported. Rassinoux's recent work on multilingual text retrieval using conceptual graphs offers some insight into how deep semantic processing might be used [60]. The system, known as RECIT, was designed for the sharply limited domain of radiology reports and hospital discharge summaries from the digestive surgery department at a single trilingual (French, English and German) hospital.[48] Rassinoux developed syntactic and semantic analysis routines to produce conceptual graphs in a manually constructed conceptual schemata, but provided no detail on how these conceptual graphs might be matched. The development of techniques for approximate matching of conceptual graphs would be useful in this regard.

Kitano's 1988 paper described a Direct Memory Access Parser (DMAP) implementation, a system he calls "SMAP," using the same hardware [36]. SMAP was designed to extract concepts from multilingual sentences and use them to fill a case frame.[49] Reported parsing speeds were better than one millisecond per word for sentences of up to 10 words. Kitano did not, however, discuss how the case frames would be designed (except to observe the need for development tools), nor did he describe how case frames would be matched.

[47] These are nine point averages, evenly spaced between 0.1 and 0.9.

[48] RECIT stands for REpreséntation du Contenu Informationnel des Textes médicaux.

[49] A multilingual sentence is one in which words from more than one language appear.

3.3 Corpus-Based Techniques

The alternative to use of a thesaurus is to directly exploit statistical information about term usage that can be gleaned from parallel corpora. This more direct approach is well suited for integration with text retrieval techniques that are themselves based on the statistics of term usage. Statistical retrieval techniques typically exploit two key observations about term usage. The first is that documents which a user would judge to be similar generally use similar terms. Referring again to the model in figure 2, q and d are typically designed to extract term frequency information and c is designed to exploit it. The second observation is that the usefulness of a term for discriminating between documents is greatest for the rarest terms and least for the most common terms. Common terms with little relation to content are typically removed by a "stoplist" of closed class terms, and the remainder are often weighted using the "inverse document frequency," which is typically calculated as:

$$idf_i = \log_2(\frac{\text{Number of documents}}{\text{Number of documents with term } i})$$

Combining the two results in the so-called "*tfidf*" (term frequency and inverse document frequency)

$$tfidf_{ij} = tf_{ij} * idf_i$$

Where tf_{ij} is the number of times term i appears in document j. More complex functions of term and document frequency are often used, so our discussion of *tfidf* is meant to be illustrative rather than exhaustive.

Several techniques can be used to construct the comparison function c for a *tfidf* representation. Probably the simplest is the "vector space technique" in which vectors of *tfidf* weights are formed by q and d and the normalized inner product of two vectors is computed by c. The normalized inner product has the desirable property that it is a strictly increasing function with respect to any decrease in the difference of two matched (i.e. same term) *tfidf* values. In other words, bringing two vectors closer along *any* dimension will increase their computed similarity. Because the normalized inner product of two vectors is the cosine of the angle between the vectors in a vector space, the normalized inner product is known as the cosine similarity measure. SMART, developed by Salton at Cornell University, is an example of a vector space text retrieval system [65].

Probabilistic retrieval techniques typically implement a more complex c function. Typically based on the simplifying assumption that j (and hence r) is binary valued (i.e., every document is either relevant or it is not), probabilistic text retrieval techniques seek to estimate the probability that a given document is relevant based on *tfidf* (or similar) evidence. INQUIRY, developed by Croft and others at the University of Massachusetts at Amherst, is an example of a probabilistic text retrieval system [14].

3.3.1 Automatic Thesaurus Construction

In a sense, corpus-based techniques can be viewed as a type of automatic thesaurus construction technique in which information about the relationship between terms is obtained from observed statistics of term usage. The difference is that in this case the "thesaurus" need not be constructed by humans. As with many other multilingual text retrieval techniques, automatic thesaurus construction has a significant research heritage in a monolingual context [15]. A substantial amount of research has appeared on this subject has been reported in the machine translation literature. For the present survey we describe two techniques for automatically constructing multilingual thesauri from a text retrieval perspective.

The first technique, developed by van der Eijk of Digital Equipment Corporation in the Netherlands, was tested on 1,100 noun phrases drawn from a parallel corpus of about 1000 long Dutch and English sentence pairs in a technical document [74].[50] The noun phrases in each sentence pair were identified using a statistical part of speech tagger and a simple parser. Candidate translations for each Dutch noun phrase were constructed by comparing the frequency with which each English term occurred in the English portion of sentence pairs containing that noun phrase to the frequency with which that English term occurred in the entire collection. An additional feature was incorporated to discourage the choice of noun phrases which occurred at significantly different relative positions in the sentence pairs.

Parameters were found that resulted in identification of the single correct translation 45% of the time, and alternative choices which produced a list of candidate translations containing the correct single translation 66% of the time were also identified. Sentence alignment, part of speech tagging and parsing errors accounted for 85% of the errors, so van der Eijk speculated that selection of the upper bound on the performance of his technique was a correct single translation about 60% of the time or inclusion of the correct translation in a list about 95% of the time. Because of the small size of the parallel corpus it was not possible to determine the performance of the technique when more than one translation of the same term was present in the corpus.[51] The resulting bilingual lexicon was not used for text retrieval, so we are unable to determine what effect the translation errors would have on retrieval effectiveness. Furthermore, we can offer no guidance regarding whether the precision reduction resulting from increasing the number of candidate translations could be offset by the recall increase resulting from a greater likelihood of including the correct translation in the list.

Recently, Lin and Chen at the University of Arizona have applied a machine learning approach to multilingual thesaurus construction [31]. Extending earlier work on term clustering, they developed a Chinese-English concept list using a collection of 1052 titles from Chinese technical papers, many of which contained a mixture of Chinese and English words. Using synaptic weights based on the pairwise co-occurrence of terms in the same title, they constructed a Hopfield neural network to generate clusters of terms.[52] Their system clustered terms from 68% of the documents into 36 concepts (without overlap), and they report that manual inspection showed that the terms associated with "all concept descriptors appeared to be relevant and precise" and that some clusters contained both Chinese and English terms. Lin and Chen also suggest that the raw term co-occurrence values could be used directly in a manner similar to the "related term" information in a conventional subject thesaurus. They report no experimental retrieval results, however.

[50] The average sentence length was over 24 words. The sentences were aligned using statistical techniques, and 7% of the sentence pairs were later discovered to be incorrectly aligned.

[51] 71% of the Dutch noun phrases occurred only once in the entire collection.

[52] In Chinese multiple symbols were recognized as phrases, but in English individual words were used.

3.3.2 Term Vector Translation

We now turn our attention to corpus-based multilingual text retrieval techniques which produce mappings that are not designed for human use. In particular, we consider statistical multilingual text retrieval techniques in which the goal is to map statistical information about term use between languages. In particular, we consider techniques which map sets of *tfidf* term weights from one language to another, a process we call *term vector translation*.

Fluhr describes a particularly simple technique which provides a good starting point for our discussion [25]. Consider a two language case in which we have three subcollections, one in English, one in French and one which is parallel (i.e. every document in the parallel collection appears in paired English and French versions). Each query is first presented to the parallel collection, and the documents in that collection are ranked with respect to the similarity between the query and the version of the documents that are in the query's language. The highest ranking French documents are then concatenated and used as a query on the remaining French documents, a variation on a technique known as relevance feedback. The same is done for the English documents. The three ranked lists are then combined in some manner and presented to the user.[53]

Relevance feedback is a commonly used technique in statistical information retrieval. A normalized *tfidf* vector is, in a sense, a heuristic approximation to the empirical distribution of term importance within a document. Viewed in this light, the normalized inner product is simply the correlation between two documents described by such distributions.[54] Since the quality of an empirical distribution can be improved by adding observations, relevance feedback can be viewed as a heuristic approach to smoothing out the clumpy empirical distributions that are associated with relatively short queries.[55] In other words, relatively unimportant terms are suppressed and relatively important terms are reinforced.

In their TREC-4 experiment, Davis and Dunning tried three more complex term vector translation techniques[16, 17, 18]. Using 80,000 pairs of aligned sentences from a parallel corpus of United Nations documents, they first selected the 8,000 English sentences that were most similar to their English translations of each TREC query. They then used the Spanish versions of those 8,000 sentences to select 100 common Spanish terms associated with each query.[56] Terms were then adaptively deleted from this set using an evolutionary programming strategy, with a goal of finding a Spanish query that could select Spanish sentences in a way similar to the way the English query selected English documents.[57] Details of the technique are presented in [17]. The evolutionary programming step only increased average precision from 0.004 to

[53]We are not aware of experimental results which describe the effectiveness of this technique.

[54]By linearity, the normalized inner product is the inner product of the normalized *tfidf* vectors.

[55]Proving such an claim would require statistical independence of the observations, a condition that is unlikely to be be satisfied. But relevance feedback has been observed to improve effectiveness, so we seek here to explain, not to prove, its effectiveness.

[56]The 100 terms chosen were those were the 501^{st} to the 600^{th} most common terms.

[57]More precisely, a Spanish query was sought which would maximize the *unnormalized* inner product of two 80,000-element vectors, one formed by computing the cosine similarity between that Spanish query and each Spanish sentence and the other formed by computing the cosine similarity between the fixed English query and each English sentence.

0.02,[58] but they observed that additional improvement might be obtained if a parallel training corpus from a domain more closely related to the evaluation domain were available.

Their third technique was based on the same training corpus of aligned sentences. Davis and Dunning chose the 100 terms with the greatest statistical significance[59] from the set of terms appearing in the Spanish sentences that were aligned with the 100 sentences most similar[60] to each English query. This technique achieved an average precision of 0.02.

Davis and Dunning's final technique was based on direct translation of term vectors [16, 23] using a linear operator. They began by forming one matrix from a collection of *tfidf* vectors derived from the English version of the aligned sentences and a second matrix derived from the Spanish versions of the same sentences. They then solved the resulting underdetermined (and potentially inconsistent) set of vector equations to find a linear operator which translated the Spanish matrix into the English one. They then used that operator to translate each English query's *tfidf* vector into a Spanish *tfidf* vector and used the translated vector to rank the Spanish documents. Davis and Dunning achieved an average precision of 0.01 using this technique. They cautioned, however, that their algorithms for computing the linear operator were still quite preliminary, so much better performance might be possible using this technique.

Oard and others at the University of Maryland have proposed another term vector translation approach based on parallel corpora which have been aligned to the word level [20, 47]. Building on term alignment techniques similar to those used by van der Eijk, they described a technique for using a bilingual term list in which alternative translations of each term are assigned (unconditioned) probability values.[61] They proposed to use this bilingual lexicon as a linear operator to map query vectors into another language.[62] They claimed that this approach would be well matched with the capability of a statistical text retrieval system to exploit imprecise information, but the technique has not yet been implemented and construction of the required bilingual term list may be a formidable task.

3.3.3 Latent Semantic Indexing

Another statistical technique that has been applied to multilingual text retrieval is Latent Semantic Indexing (LSI) [21]. The basic idea is to use a matrix decomposition to identify the principal components of the vector space defined by the document collection, and then project the vectors into the space spanned by those principal components. In LSI the principal components are thought to represent important conceptual distinctions, while the lesser components are thought to represent term usage variations. So LSI seeks to emphasize the important aspects of the *tfidf* distribution

[58] Recall that they achieved an average precision of 0.04 with unconstrained query expansion.

[59] The statistical significance of each term was estimated using a likelihood ratio test, comparing term frequency in the selected set with term frequency in the entire collection.

[60] Again, similarity was computed using the cosine measure.

[61] Oard, et. al. actually cited the work of Brown, et. al. [8] on statistical machine translation. Those techniques are more complex than van der Eijk's, but they have been applied to much larger parallel corpora.

[62] They proposed this technique in the context of vector space text filtering, in which a vector representation of the information need is constructed without reference to an explicit query.

and suppress the effect of varying term usage. Documents can then be compared using the cosine similarity measure and ranked for display in the usual way.

LSI has been applied to multilingual text retrieval in a similar way to the relevance feedback technique described above [5, 37, 38, 76]. The basic approach is best illustrated by Landauer and Littman [38]. Randomly selecting 900 training paragraphs and 1,582 evaluation paragraphs from the Hansards collection, a parallel corpus of Canadian parliamentary proceedings, they first applied LSI to identify the principal components of the training set. When LSI is applied to a parallel corpus, the matrix decomposition naturally identifies the principal components in the vector space associated with each language and produces a mapping from each to a common representation space with fewer dimensions. They then selected the principal components of the *tfidf* vector for every paragraph in the evaluation set, regardless of language, in this common representation space. Using the English vectors as queries, they found that the top ranked French vector was derived from the translated version of the English paragraph in 92% of the 1,582 cases. Unfortunately, the lack of a bilingual corpus with available relevance judgements precluded a more traditional recall-precision evaluation.

Berry and Young repeated this work using passages from the Bible in English and Greek [5]. They were able to demonstrate that fine-grained training data, using only the first verse of each passage to identify the principal components, improved retrieval performance over Landauer and Littman's coarse-grained approach. Using 16 short queries, each of which had between two and six relevant passages in a collection of 734 passages which they constructed.[63] Rather than report precision-recall results they observed that the average rank of a relevant document decreased from about sixth to fourth when the same number of training verses were distributed across every passage in the collection rather than clustered in a small group of passages.

In an interesting combination of corpus-based and thesaurus-based techniques, Evans and others at Carnegie Mellon University used LSI to suggest terms from a controlled vocabulary of 125 English medical terms based on natural language queries expressed in Spanish [24]. Augmenting definitions found in three English medical thesauri with related words from both English and Spanish, they obtained a training set of 3,084 words.[64] Their report presents two examples in which the most highly ranked terms would be good choices for use in a controlled vocabulary search.

3.4 Other research projects

In addition to the work we have cited, we are aware of three other research groups investigating multilingual text retrieval. Because we know of no published research results from these projects, we simply describe their stated objectives briefly.

In December of 1993 a team led by Laus-Maczynska of the French firm Cap Gemini Innovation began work on the CRISTAL project [11].[65] A part of the I*M-Europe Language Engineering program, CRISTAL is being designed to retrieve documents from a French collection using queries in French, English or Italian using the French Dicologique thesaurus. It is scheduled for completion in May of 1996.

[63] The queries contained between one and four words.

[64] Evans, et. al., used term definitions from the QMR, PTXT and UMLS META-1 thesauri.

[65] CRISTAL stands for Conceptual Retrieval of Information using Semantic dicTionAry in three Languages.

Liddy, of Syracuse University and Textwise Inc., began a feasibility study of multilingual text retrieval for the Advanced Research Projects Agency (ARPA) in 1994 [1]. The proposed system, known as CINDOR, was designed to exploit a multilingual thesaurus for concept retrieval.

Finally, in a 1995 research paper on adaptation of the INQUIRY probabilistic text retrieval system for monolingual text retrieval in languages other than English, Croft and others from the University of Massachusetts briefly described future plans to investigate cross-linguistic text retrieval [14].

4 Some Observations on the State of the Art

We will now take advantage of the background we have developed to make a few observations on the present state of multilingual text retrieval practice and research.

Controlled vocabulary techniques are extremely well developed, but fully automatic thesaurus construction is still in its infancy. Furthermore, multilingual concept retrieval techniques such as query expansion that could exploit information encoded in a thesaurus without human intervention at indexing or retrieval time have thus far been limited to approximating the within-language effectiveness of the *same* technique in the same domain. Without effective automatic thesaurus construction, the limited domain of concept retrieval techniques will remain a serious limitation.

The relative immaturity of corpus-based techniques means that thesauri are presently an important component of any practical multilingual text retrieval system, regardless of whether an exact match or a ranked retrieval model is adopted. Furthermore, integration of thesauri with techniques based on corpus statistics is an area of active research in computational linguistics, and there is some indication that the best features from each can be captured when the two techniques are combined [30]. Because the most sophisticated multilingual text retrieval thesauri in existence are in controlled vocabulary systems, ongoing research efforts would likely benefit from leveraging what has been learned in this work.

The differing domains of available parallel corpora and scored corpora (corpora for which relevance judgements are available) remains the largest single obstacle to evaluation of corpus-based techniques. We are not aware of a single instance of a large parallel corpus with an associated set of queries for which relevance judgements are available. Without such a corpus, the best possible experiment design is to train on a parallel corpus from a domain similar to that of the evaluation corpus. Unfortunately, we are not aware of any techniques for estimating the effects, or even the degree of a mismatch between the training and the evaluation domain. Without either scored parallel corpora or some way of estimating the effect of a domain difference it will not be possible to draw conclusive observations from large-scale studies such as those conducted by Davis and Dunning [16].

The performance of monolingual techniques under the same experimental conditions appears to be a good benchmark for an upper bound on retrieval effectiveness. There is presently no evidence that multilingual techniques can reliably exceed the performance of monolingual techniques. Fluhr and Radwan have demonstrated that it is reasonable to lower bound the effectiveness of a multilingual text retrieval system with the effectiveness of a modular approach in which fully automatic machine translation to preprocess the query, and our analysis in section 3 supports this assertion.

Agreement on these two common points of reference would facilitate comparison of multilingual text retrieval approaches across different experiments. The resources required to realize the potential of modern fully automatic machine translation systems may limit the utility of this approach in smaller studies, however.

One important difference between monolingual and multilingual retrieval is that polysemy appears to be a key limiting factor. In particular, polysemy seems to become a problem more rapidly in multilingual retrieval than in monolingual retrieval as the size of the domain increases. Three researchers, operating with very different experiment designs, have confirmed that polysemy can be reduced using syntactic and semantic information, of which the simplest type is phrase formation. This suggests that word sense disambiguation which, like phrase formation, has demonstrated limited utility in a monolingual context might be a productive avenue for further investigation.

The key issue in application of any natural language processing technique to multilingual text retrieval is improving precision without a significant adverse effect on recall. This argues for investigating relatively shallow techniques that can be designed to degrade gracefully as the domain drifts. One of the pitfalls of translating queries is that short queries may increase the adverse effect of polysemy by limiting contextual clues about word sense. In order to deal with this effect, Hull and Grefenstette have proposed using structural information from the document space to enhance domain-specific interpretation of the query [32] and Radwan and Fluhr have implemented a simple version of this approach. In contrast, Oard and his colleagues suggested exploiting the structure of user interest evidence gained over time [47]. The two approaches seem complementary, with the decision between them depending on the relative rate at which the document space and the users' information needs are changing.

5 Conclusion

In use since 1965, controlled vocabulary multilingual text retrieval systems are clearly able to provide satisfactory solutions in some applications. That fact often seems to be overlooked, however, by researchers seeking to develop techniques suitable for cost-effective application in broad domains. This appears to reflect a dichotomous world view between Library Science, which has embraced exact match concept retrieval, and Computer Science, which has embraced natural language ranked retrieval systems. Because they fill different niches, the two disciplines have developed conflicting terminology for similar concepts. One goal of this survey has been to unify those two world views.

We have described a taxonomy of multilingual text retrieval approaches that is based on a fundamental division into thesaurus-based and corpus-based approaches. Controlled vocabulary and concept retrieval are the two dominant approaches to thesaurus utilization. Deeper semantic processing has been applied in a few cases, most notably in the EMIR project. Automatic thesaurus construction bridges the gap between the thesaurus-based and corpus-based approaches, and the linear and nonlinear approaches to term vector translation complete the taxonomy. We have contrasted these approaches with a modular "translate-then-retrieve," explaining how a more integrated approach can achieve better performance with less effort. The experimental results obtained in the EMIR project agree with our assessment.

Two important issues that deserve increased attention from the research community have been identified. One is the lack of a large scored multilingual corpus or, failing that, some principled way of interpreting the results of corpus-based experiments in which the training corpus and the evaluation corpus address different domains. The other issue is how to mitigate the adverse effects of polysemy on cross-language retrieval effectiveness. Although this issue has been studied extensively in a monolingual context, it appears that a critical reevaluation of the available techniques in the multilingual context could be quite productive.

As improved communication increases the interdependence between nations, multilingual text retrieval will become an increasingly important technology. The controlled vocabulary approach used by existing systems will undoubtedly continue to be used in applications where it's strengths can be exploited. But new techniques will be needed as well, and the research we have described provides a basis on which to develop those techniques.

Acknowledgement

The authors would like to express their appreciation to Dagobert Soergel, David Hull, Natalie Schoch and Gary Marchionini for their insightful comments and to Nicholas DeClaris and for sustained support for this research.

References

[1] FY 1994 SBIR solicitation, Phase I award abstracts, ARPA projects. Defense Technical Information Center, 8725 John J. Kingman Road, Suite 0944, Ft. Belvior, VA 22060, 1994. ftp://ftp.dtic.dla.mil/pub/sbir/arpa94sbir_awds.

[2] Belal Mustafa Abu Ata, Tengku Mohd. Tengku Sembok, and Mohammed Yusoff. Sisdom: a multilingual document retrieval system. *Asian Libraries*, 4(3):37–46, September 1995.

[3] Derek Austin. Progress towards standard guidelines for the construction of multilingual thesauri. In Commission on the European Communities, editor, *Third European Congress on Information Systems and Networks*, volume 1, pages 341–402. Verlag Dokumentation, May 1977.

[4] Heiner Benking and Ulrich Kampffmeyer. Harmonization of environmental meta-information with a thesaurus-based multi-lingual and multi-medial information system. In Arthur Zygielbaum, editor, *AIP Conference Proceedings 283, Earth and Space Science Information Systems*, pages 688–695. American Institute of Physics, 1992.

[5] M. Berry and P. Young. Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities*, 29(6):413–429, December 1995.

[6] Paul Blake. The MenUSE system for multilingual assisted access to online databases. *Online Review*, 16(3):139–145, 1992.

[7] P. Bollmann and E. Konrad. Automatic association methods in the construction of interlingual thesauri. In W. E. Batten, editor, *EURIM II A European Conference*

on the Application of Research in Information Science and Libraries, pages 152–155. Aslib, 1976.

[8] Peter F. Brown, John Cocke, Steven A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June 1990.

[9] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using SMART: TREC 3. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 69–80. NIST, November 1994. http://potomac.ncsl.nist.gov/TREC/trec3.papers/cornell.new.ps.

[10] C. Cacares. Russian-Spanish multisubject computer dictionary. *Automatic Documentation and Mathematical Linguistics*, 20(2):122–125, 1986. English translation from Russian.

[11] Roberto Cencioni and Ewan Klein. Telematics programme 1991-1994 Language Research & Engineering (LRE) an overvieww. Directorate General XIII, Commission of the European Communities, June 1994. http://www2.echo.lu/LangEng/en/lrefacts.html.

[12] Vinod Chachra. Subject access in an automated multithesaurus and multilingual environment. In Sally McCallum and Monica Ertel, editors, *Automated Systems for Access to Multilingual and Multiscript Library Materials*, pages 63–76. International Federation of Library Associations and Institutions (IFLA), K. G. Saur, August 1993.

[13] Ewa Chmielewska-Gorczyca and Waclaw Struk. Translating multilingual thesauri. In Pavla Stančiková and Ingetraut Dahlberg, editors, *Proceedings of the First European ISKO Conference*, pages 150–155. International Society for Knowledge Organization, Indeks Verlag, September 1994.

[14] W. B. Croft, J. Broglio, and H. Fujii. Applications of multilingual text retrieval. In *Proceedings of the Twenty-Ninth Annual Hawaii International Conference on System Sciences*, pages 98–107, 1995.

[15] C. J. Crouch. An approach to the automatic construction of global thesauri. *Information Processing and Management*, 26(5):629–640, 1990.

[16] Mark Davis and Ted Dunning. A TREC evaluation of query translation methods for multi-lingual text retrieval. In D. K. Harman, editor, *The Fourth Text Retrieval Conference (TREC-4)*. NIST, November 1995. http://crl.nmsu.edu/ANG/MWD/Book2/trec4.ps.

[17] Mark W. Davis and Ted E. Dunning. Query translation using evolutionary programming for multi-lingual information retrieval. In *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, March 1995. http://crl.nmsu.edu/ANG/MWD/Book2/evolmltr1.ps.gz.

[18] Mark W. Davis and Ted E. Dunning. Query translation using evolutionary programming for multilingual information retrieval II. In *Proceedings of the Fifth Conference on Evolutionary Programming*, March 1996. To Appear. http://crl.nmsu.edu/ANG/MWD/Book2/ep96.ps.

[19] Carmen Lopez de Sosoaga. Multilingual access to documentary database. In A. Lichnerowicz, editor, *Proceedings of a Conference on Intelligent Text and Image Handling (RIAO 91)*, pages 774–788, Amsterdam, April 1991. Elsevier.

[20] Nicholas DeClaris, Donna Harman, Christos Faloutsos, Susan Dumais, and Douglas Oard. Information filtering and retrieval: Overview, issues and directions. In Jr. Norman F. Sheppard, Mary Eden, and Gideon Knator, editors, *Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 42a–49a. IEEE, November 1994. http://www.ee.umd.edu/medlab/filter/papers/balt.ps.

[21] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. http://superbook.bellcore.com/ std/papers/JASIS90.ps.

[22] C. P. R. Dubois. Free text vs. controlled vocabulary; a reassessment. *Online Review*, 11(4):243–253, August 1987.

[23] Ted E. Dunning and Mark W. Davis. Multi-lingual information retrieval. Memoranda in Cognitive and Computer Science MCCS-93-252, New Mexico State University, Computing Research Laboratory, February 1993. http://crl.nmsu.edu/ANG/MWD/Book2/mltr.ps.gz.

[24] D. A. Evans, S. K. Handerson, I. A. Monarch, J. Pereiro, L. Delon, and W. R. Hersh. Mapping vocabularies using "latent semantics". Technical Report CMU-LCL-91-1, Carnegie Mellon University, Laboratory for Computational Linguistics, July 1991.

[25] Christian Fluhr. Multilingual information retrieval. In Ronald A Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology*, pages 391–305. Center for Spoken Language Understanding, Oregon Graduate Institute, 1995. http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node7.html.

[26] Christian Fluhr and Khaled Radwan. Fulltext databases as lexical semantic knowledge for multilingual interrogation and machine translation. In Patrick Brezillon and Vadim Stefanuk, editors, *Proceedings of the East-West Conference on Artificial Intelligence (EWAIC '93)*, pages 124–128, Moscow, September 1993. Association for Artificial Intelligence of Russia, ICSTI.

[27] Office for Official Publications of the European Communities. *Thesaurus EUROVOC Volume 3: Multilingual version.* Luxembourg, 1995.

[28] William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, NJ, 1992.

[29] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the Association for Computing Machinery*, 30(11):964–971, November 1987.

[30] Louise Guthrie, James Pustejovsky, Yorick Wilks, and Brian M. Slator. The role of lexicons in natural language processing. *Communications of the Association for Computing Machinery*, 39(1):63–72, January 1996.

[31] Chung hsin Lin and Hsinchun Chen. An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. *IEEE Transactions on Systems, Man and Cybernetics*, 26(1):75–88, February 1996. http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html.

[32] David A. Hull and Gregory Grefenstette. Experiments in multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996. To appear. http://www.xerox.fr/grenoble/mltt/people/hull/papers/sigir96.ps.

[33] A. Iljon. Creation of thesauri for EURONET. In Commission of the European Communities, editor, *Third European Congress on Information Systems and Networks*, volume 1, pages 417–437. Verlag Dokumentation, May 1977.

[34] Ariane Iljon. Scientific and technical data bases in a multilingual society. *On-Line Review*, 1(2):133–136, 1977.

[35] S. Ya. Kalachkina. Algorithmic determination of descriptor equivalents in different natural languages. *Automatic Documentation and Mathematical Linguistics*, 21(4):21–29, 1987. English translation from Russian.

[36] Hiroaki Kitano. Multilingual information retrieval mechanism using VLSI. In A. Lichnerowicz, editor, *RIAO 88 Program: User-Oriented Content-Based Text and Image Handling*, volume 2, pages 1044–1059, March 1988.

[37] Thomas K. Landauer and Michael L. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38. UW Centre for the New OED and Text Research, Waterloo Ontario, October 1990. http://www.cs.brown.edu/people/mll/x-lang.ps.

[38] Thomas K. Landauer and Michael L. Littman. A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the Eleventh International Conference: Expert Systems and Their Applications*, volume 8, pages 77–85. Avignon France, May 1991.

[39] Abraham I Lebowitz, Robert Portegies Zwart, and Helga Schmid. Multilingual indexing and retrieval in bibliographic systems: The AGRIS experience. *Quarterly Bulletin of the International Association of Agricultural Librarians and Documentalists*, 36(3):187–192, 1991.

[40] C. S. Li, A. S. Pollitt, and M. P. Smith. Multilingual MenUSE - a Japanese front-end for searching English language databases and vice versa. In *Proceedings of the 14th BCS IRSG Research Colloquium on Information Retrieval*. Springer-Verlag, April 1992.

[41] B. R. Loginov and V. V. V'yugin. Automated maintenance of a bilingual medical thesaurus on a microcomputer. *Automatic Documentation and Mathematical Linguistics*, 23(2):72–75, 1989. English translation from Russian.

[42] Richard S. Marcus. Intelligent assistance for document retrieval based on contextual, structural, interactive Boolean models. In *RIAO 94 Conference Proceedings, Intelligent Multimedia Information Retrieval Systems and Management*, volume 2, pages 27–43, Paris, October 1994. Centre de Hautes Etudes Internationales d'Informatique Documentaire (C.I.D.).

[43] P. Nelson. Breaching the language barrier: Experimentation with Japanese to English machine translation. In David I Raitt, editor, *15th International Online Information Meeting Proceedings*, pages 21–33. Learned Information, December 1991.

[44] H. Neville. Session V report of the English language discussion group. In *Second European Congress on Information Systems and Networks*, pages 162–164. Verlag Dokumentation, May 1975.

[45] H. H. Neville. Feasibility study of a scheme for reconciling thesauri covering a common subject. *Journal of Documentation*, 26(4):313–336, December 1970.

[46] H. H. Neville. Alternatives to conventional multilingual thesauri. In Verina Horsnell, editor, *Report of a Workshop on Multilingual Systems*, pages 10–12, 1975. British Library Research and Development Report 5265 HC.

[47] Douglas W. Oard, Nicholas DeClaris, Bonnie J. Dorr, and Christos Faloutsos. On automatic filtering of multilingual texts. In *Conference Proceedings, 1994 IEEE International Conference on Systems, Man, and Cybernetics*, volume 2, pages 1645–1650, October 1994. http://www.ee.umd.edu/medlab/filter/papers/smc.ps.

[48] Irma Pasanen-Tuomainen. Analysis of subject searching in the TENTTU books database. In Jay K. Lucker, editor, *Proceedings of the 14th Biennial Conference of IATUL*, volume 1, pages 72–77. International Association of Technological University Libraries, June 1991.

[49] N. A. Pashcenko, S. Ya. Kalachkina, N. M. Matsak, and V. A. Pigur. Basic principles for creating multilanguage information retrieval thesauri (experience with implementing GOST 7.24-80). *Automatic Documentation and Mathematical Linguistics*, 16(3):30–36, 1982. English translation from Russian.

[50] D. Pelissier and O. Artur. The multilingual evolution of PASCAL. In *10th International Online Information Meeting*, pages 113–121. Learned Information, December 1986.

[51] B. R. Pevzner. Automatic translation of English text to the language of the Pusto-Nepusto-2 system. *Automatic Documentation and Mathematical Linguistics*, 3(4):40–48, 1969. English translation from Russian.

[52] B. R. Pevzner. Comparitive evaluation of the operation of the Russian and English variants of the "Pusto-Nepusto-2" system. *Automatic Documentation and Mathematical Linguistics*, 6(2):71–74, 1972. English translation from Russian.

[53] V. A. Pigur. Multilanguage information-retrieval systems: Integration levels and language support. *Automatic Documentation and Mathematical Linguistics*, 13(1):36–46, 1979. English translation from Russian.

[54] A. Steven Pollitt and Geoff Ellis. Multilingual access to document databases. In *21st Annual Conference Canadian Society for Information Science*, pages 128–140, July 1993.

[55] A. Steven Pollitt, Geoffrey P. Ellis, Martin P. Smith, Mark R. Gregory, Chun Sheng Li, and Henrik Zangenberg. A common query interface for multilingual document retrieval from databases of the European Community institutions. In *Proceedings of the 17th International Online Information Meeting*, pages 47–61, December 1993.

[56] Barbara Rad-El. Approaches to multilanguage and multiscript issues in the ALEPH system. In Sally McCallum and Monica Ertel, editors, *Automated Systems for Access to Multilingual and Multiscript Library Materials*, pages 145–150. International Federation of Library Associations and Institutions, K. G. Saur, August 1993.

[57] Khaled Radwan. *Vers l'Accés Multilingue en Langage Naturel aux Bases de Données Textuelles*. PhD thesis, Université de Paris-Sud, Centre d'Orsay, 1994.

[58] Khaled Radwan and Christian Fluhr. Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 121–136, April 1995.

[59] Khaled Radwan, Frederic Foussier, and Christian Fluhr. Multilingual access to textual databases. In A. Lichnerowicz, editor, *Proceedings of a Conference on Intelligent Text and Image Handling (RIAO 91)*, pages 475–489. Elsevier, April 1991.

[60] A. M. Rassinoux, R. H. Baud, and J. R. Scherrer. A multilingual analyser of medical texts. In W. M. Tepfenhart, J. P. Dick, and J. F. Sowa, editors, *Second International Conference on Conceptual Structures (ICCS 94)*, pages 84–96. Springer-Verlag, 1994. http://www.hbroussais.fr/helios/doc/nlp/Rassinoux94b.html.

[61] Paule Rolland-Thomas and Gérard Mercure. Subject access in a bilingual online catalog. *Cataloging and Classification Quarterly*, 10(1/2):141–163, 1989.

[62] Loll Rolling. Multilingual systems: survey of the European scene. In Verina Horsnell, editor, *Report of a Workshop on Multilingual Systems*, pages 4–5, October 1975. British Library Research and Development Report 5265 HC.

[63] G. Salton. Experiments in multi-lingual information retrieval. *Information Processing Letters*, 2(1):6–11, 1973. TR 72-154 at http://cs-tr.cs.cornell.edu.

[64] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194, May 1970.

[65] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[66] F. Semturs. Information retrieval from documents in multilingual textual data banks. In *Third European Congress on Information Systems and Networks*, pages 463–467, Munich, May 1977. Verlag Dokumentation.

[67] Fritz Semturs. STAIRS/TLS - a system for "free text" and "descriptor" searching. In Everett H. Brenner, editor, *Proceedings of the ASIS Annual Meeting*, volume 15, pages 295–298. American Society for Information Science, November 1978.

[68] M. P. Smith, A. S. Pollitt, and C. S. Li. An evaluation of concept translation through menu navigation in the MenUSE intermediary system. In Tony McEnery and Chris Paice, editors, *14th Information Retrieval Colloquium*, pages 38–54. British Computer Society, Springer-Verlag, April 1992.

[69] Dagobert Soergel. Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science*, 45(8):589–599, September 1994.

[70] Erwin Stegentritt. *German Analysis: Morpho-Syntax Within the Framework of the Free-Text Retrieval Project E.M.I.R.*, volume 15. AQ-Verlag, Saarbrucken, Germany, 1994.

[71] Catherine Synellis. TRANSLIB user survey report. TRANSLIB technical report, University of Patras Central Library, Rio 261 00 Patras, Greece, May 1995. http://grial.uc3m.es/ aedo/translib/UserAn.htm.

[72] M. Tallving and P. Nelson. Japanese databases and machine translation: A question of international accessibility to Japanese databases. In David I Raitt, editor, *14th International Online Information Meeting Proceedings*, pages 423–437. Oxford, Learned Information, December 1990.

[73] United Nations Educational, Scientific and Cultural Organization (UNESCO). Guidelines for establishment and development of multilingual scientific and technical thesauri for information retrieval. Place de Fontenoy, Paris 7e, December 1971. SC/WS/501.

[74] Pim van der Eijk. Automating the acquisition of bilingual terminology. In *Sixth Conference of the Europeran Chapter of the Association for Computational Linguistics*, pages 113–119, April 1993.

[75] K. I. Volodin, L. L. Gul'nitskii, R. N. Maksakova, V. F. Parkhomenko, I. F. Pozhariskii, L. V. Fedotova, and N.I. Yakovleva. Bilingual indexing of geological documents. *Automatic Documentation and Mathematical Linguistics*, 25(6):43–45, 1991. English translation from Russian.

[76] Paul G. Young. Cross-language information retrieval using latent semantic indexing. Technical Report CS-94-259, University of Tennessee, Knoxville, December 1994. http://www.cs.utk.edu/ library/TechReports/1994/ut-cs-94-259.ps.Z.

A note on the references

The breadth and variety of the work on multilingual text retrieval would make production of a comprehensive bibliography on the subject an enormous undertaking. However, within the scope of the survey we believe that our bibliography is representative, covering every significant technique and providing citations to every important research project of which we are aware. Our survey was generally restricted to documents in the English language, however, and only exceptional work in other languages has been cited. Where Uniform Resource Locators (URL) are included in the citation, they were believed to be correct at the time of publication but may have changed since. Current links to to every online multilingual text retrieval reference of which we are aware (including those in languages other than English) can be found at http://www.ee.umd.edu/medlab/mlir/. The first author would appreciate being notified of additional online resources or changed URL's by electronic mail.