

Survey: Multilingual Text Retrieval and Access

Hachim Haddouti

FORWISS (Bavarian Research Center for Knowledge-Based Systems)
Orleansstr. 34
80689 Munich, Germany
haddouti@forwiss.de

February 1999

Table of Content

1. Introduction	3
2. Multilingual Text Retrieval.....	4
2.1 Transliteration/Transcription	4
2.2 Controlled Vocabulary	5
2.3 Knowledge-based Technique	6
2.4 Related Projects.....	8
3. Multilingual Text Access	11
3.1 Character Set	12
3.2 User Interface	13
3.3 Internationalization Of WWW (HTML, URL, HTTP)	14
3.4 Related Activities	15
4. Conclusion	19
5. Internet Resources	19
5.1 Working Groups and Related Activities.....	19
5.2 Tools and Demos.....	22
5.3 Corpora, Lexicons, Dictionaries etc.	22
5.4 Selected Publications and Related Resources	24
6. References	25
Glossary	27
Acronyms	28

1. Introduction

1 Although 6,700 languages are spoken in 228 countries and English is the native language of only 6 % of the World population, English is the dominant language of the collections, resources and services in the Internet. Actually, the English language is widely used on the Internet. About 60 % of the world online population is represented by English, and 30% by European languages [NUA98] (October 98)¹. Approximately 147 M people are connected to the Internet (US and Canada about 87 M, Europe 33.25 M, Asia/Pacific 22 M, Africa 800,000). However, the size of web sites and Internet users from other countries (non-English countries) is increasing progressively, so that the multilingual products will soon reach their high level of importance.

2 The time of globalization is coming. Many countries have been unified. The European project to unify European countries is a very important example in order to eliminate broader for the cooperation, global and large market, real international and free business. The high-developed technologies in network infrastructure and Internet set the platform of the cooperation and globalization. Indeed, the business should be global and worldwide oriented. Thus, the issues of the multilinguality arise and should be addressed as soon as possible in order to overcome the remaining technical barriers that still separate countries and cultures.

3 In the near future, a lot of digital libraries will be set up containing large collections of information in a large number of languages. However, it is impractical to submit a query in each language in order to retrieve these multilingual documents. Therefore, a multilingual retrieval environment is essential for benefiting from worldwide information resources.

4 The development of digital libraries is becoming increasingly relevant. However, most research and development activities are focussed on only one language. But, this is not the objective of the digital libraries and Internet philosophy. Both technologies aim at establishing a global digital library containing all information resources from different areas, different countries and in different languages. The access to those materials should be possible for the worldwide community and never be restricted because of non-understanding languages. Thus, the EU funded some projects addressing the multilingual issues. For instance, in the ESPRIT project EMIR (European Multilingual Information Retrieval) a commercial information retrieval system SPIRIT has been developed which supports French, English, German, Dutch and Russian. UNESCO launched some projects in order to democratize and globalize the access to the world cultural patrimony, such as Memory of the World. Recently, UNESCO has started the MEDLIB² project which aims at creating a virtual library for the Mediterranean region. The Mediterranean basin presents different cultural, linguistic and historical patrimonies. The language diversity of this region will play the key role in launching projects addressing the multilingual issues. The sooner the world and in particular the Mediterranean community becomes involved in these discussions and projects, the sooner we will benefit from the high-tech development worldwide.

5 In this study we will survey the state of the art in multilingual text retrieval and access. The survey will benefit the MEDLIB project in planning a strategy how to retrieve and access the holdings of the Virtual Library MEDLIB, which will contain documents in different languages (Arabic, Hebrew, West-European language, Greek, etc.). Various techniques will be described and their weaknesses will be discussed. Moreover, developed products and related projects will be reported. Therefore it will facilitate the choice between a huge amount of products and to distinguish between commercial and public domain tools. However, the machine translation products are not the subject of this survey.

6 The rest of this paper is designed as follows. In the second Section we will survey the multilingual text retrieval technique and present selected projects and products in this field, while Section 3 gives

¹ The most recent statistics are available at <http://www.nua.net>

² <http://www.unesco.org>

us an overview about the multilingual text access and its issues. Moreover, well known projects and solutions are discussed. Section 4 concludes this survey. In Section 5, we present a collection of Internet resources, activities and working groups regarding our topic. At the end of this document a glossary and an acronym description table are given.

2. Multilingual Text Retrieval

7 The approach of cross-language information retrieval allows a user to formulate a query in one language and to retrieve documents in others. The controlled vocabulary is the first and the traditional technique widely used in libraries and documentation centers. Documents are indexed manually using fixed terms which are also used for queries. These terms can be also indexed in multiple languages and maintained in a so-called thesaurus.

8 Using dictionary-based technique queries will be translated into a language in which a document may be found. The corpus-based technique analyzes large collections of existing texts and automatically extracts the information needed on which the translation will be based. However, this technique tends to require the integration of linguistic constraints, because the use of only statistical techniques by extracting information can introduce errors and thus achieve bad performance [haddouti98].

9 The development of multilingual retrieval systems is very limited because of high cost and complexity. Most of those applications are based on thesauri which are very expensive to implement. Stemming, word boundary identification, and lists of stopwords must be defined. From language to language term indexing differs. Some languages are written without spaces between words. In this case character-based indexing is more suitable than word-based indexing. Thus, profound investigations in the following areas are necessary: machine translation systems, natural language processing, advanced linguistic processing tools, morphological analysis, lexical semantic information extraction, terminology extraction, algorithms for alignment of translated texts.

10 Several information retrieval system issues have already been studied for other languages than the English language. Examples of these languages are Chinese, Japanese, Spanish, etc. Both Spanish and Chinese retrievals have been evaluated in TREC 96 (Text Retrieval Evaluation Conference, NIST, 1996).

11 In this Section we will provide an overview of the multilingual text retrieval approaches and present in detail the developed techniques and their drawbacks.

2.1 Transliteration/Transcription

12 In many applications, documents such as geographic maps, bibliographic data, directory services, etc. written in non-Roman scripts such as Japanese, Arabic, Chinese, Hebrew, etc. are transliterated or transcribed into Roman characters. Transliteration is the process of converting the characters of an alphabetical or syllabic script to the characters of a conversion alphabet. In some cases, the use of diacritics and digraphs is required to keep the phonetic sound acceptable.

13 However, transliteration is not suitable for ideographic characters, such as CJK (Chinese Japanese and Korean), because their transliterations into alphabets are identical and therefore lose their meaning. For these languages the use of transcription is required. Transcription is the process whereby the sounds of a given language are noted by the system of signs of a conversion language.

14 Transliteration/transcription matches characters syntactically; it does not translate meaning. This technique leads to a huge loss of information. For instance, the Spanish language contains many Arabic words which have been more or less transliterated. There are some words which are unfamiliar

for the Arabic-native speakers because they sound very different from the original Arabic phonetics, e.g. *Ojalà* (hopefully), *Gibraltar* (the British colony of Gibraltar in Spain). Other types of loss are accents, umlauts and other language-specific mark variants forms of words, which do not match properly.

15 An international forum, ISO/TC46/SC2³ (International Technical Committee 46 / Subcommittee Conversion of Written Languages) has been built. Its goal is to create Standards for conversion methods of writing systems. It has produced a series of International Standards covering transliteration and/or transcription of several scripts (e.g. Arabic, Cyrillic, Greek, Hebrew, Persian, Korean, Mongolian, Thai, etc.) into Latin characters. An electronic mailing list for TC46/SC2, called tc46sc2@elot.gr, has been created to discuss the issues and research results.

2.2 Controlled Vocabulary

16 A controlled vocabulary is the widely used traditional technique applied in (multilingual) information retrieval. Terms must be assigned to all documents of the collection. To this end, descriptors are given to represent a documents. The terms are translated and mapped to each other in thesauri. However the translation of terms is not isomorphic, i.e. a term can have many translations and can have different meanings in different concepts.

17 In 1971, UNESCO proposed standards for multilingual thesaurus development [UNESCO71]. Five years later the ISO expanded the draft specification and then approved it in 1978 as ISO 5964. The European Community launched the EUROVOC project in order to build a multilingual thesaurus based on the established standard ISO 5964. The EUROVOC thesaurus supporting nine official European Community languages was published 1984.

18 In [TRANSLIB95] multilingual tools, such as bilingual dictionaries, grammar parser, machine translation tools (e.g. Systrans, Rosette), terminological databases, thesauri, etc. have been evaluated according to different criteria. This evaluation demonstrated how these tools will be used in the TRANSLIB project to allow trilingual (Greek, Spanish and English) search in the library catalogs (see also Section 2.4). Some assessments in this project and others showed that the controlled vocabulary text retrieval technique is widely used in libraries.

19 Building thesauri can be done manually or automatically. By the last one, key words of documents can be collected and classified and organized which will produce a thesaurus. Semantic relationships (synonyms, neighborhood, etc.) between terms will be deduced. In multilingual thesauri, terms of different languages are indexed and mapped to each other according to the ISO 5964 specifications. A thesaurus can be also seen as an ontology because it provides relationships between concepts and terms. For instance, the user first looks for a concept in thesaurus in order to find terms in a certain language. Those terms will be used for the search process.

20 However, this approach remains limited to application whose vocabulary is still manageable. Once the size of the vocabulary increases, the efficiency and effectiveness degrade radically. Furthermore, this approach requires that documents will be indexed using a selected vocabulary and that terms from the vocabulary must be assigned to each document in the collection [Soergel97]. The users can only use these predefined terms to formulate their queries. Here it lacks flexibility and free text search.

³ <http://www.elot.gr/tc46sc2/>

2.3 Knowledge-based Technique

Dictionary-based Approach

21 The size of public domain and commercial dictionaries in multiple languages on the Internet is increasing steadily (see Web collections in Section 5). As an example we cite a few of them: Collins COBUILD English Language Dictionary and its series in major European languages, Leo Online Dictionary, Oxford Advanced Learner's Dictionary of Current English, Webster's New Collegiate Dictionary.

22 Electronic monolingual and bilingual dictionaries build a solid platform for developing multilingual applications. Using dictionary-based technique queries will be translated into a language in which a document may be found. However, this technique sometimes achieves unsatisfactory results because of ambiguities. Many words do not have only one translation and the alternate translations have very different meanings. Moreover, the scope of a dictionary is limited. It lacks in particular a technical and topical terminology which is very crucial for a correct translation. Nevertheless, this technique can be used for implementing simple dictionary-based application or can be combined with other approaches to overcome the above mentioned drawbacks.

23 Using electronic dictionary-based approach for query translation has achieved an effectiveness of 40-60% in comparison with monolingual retrieval [Ballesteros, Croft96], [Hull, Grenfenstette96]. In [Hayashi et al. 97] a multilingual search engine, called TITAN, has been developed. Based on a bilingual dictionary it allows to translate queries from Japanese to English and English to Japanese. TITAN helps Japanese users to search in the Web using their own language. However, this system suffers again from polysemy (see Section 2.4).

Corpus-based Technique

24 The corpus-based technique seems to be promising. It analyzes large collections of existing texts (corpora) and automatically extracts the information needed on which the translation will be based. Corpora are collections of information in electronic form to support e.g. spelling and grammar checkers, and hyphenation routines (lexicographer, word extractor or parser, glossary tools).

25 These corpora are used by researchers to evaluate the performance of their solutions, such TREC collections for cross-language retrieval. A few examples of mono, bi- and multilingual corpora are Brown Corpus, Hansard and United Nation documents respectively. The Hansard Corpus⁴ contains parallel texts in English and Canadian French collected during six years by the Canadian Parliament. The Brown Corpus consists of more than one million words of American English. It was published 1961 and it is now available at the ICAME⁵ (International Computer Archive of Modern English). Interested readers are referred to the survey about electronic corpora and related resources in [Edwards 93].

26 In the United States, the WordNet Project at Princeton has created a large network of word senses in English related by semantic relations such as synonymy, part-whole, and is-a relations [Miller90] and [Fellbaum98]. Similar work has been launched in Europe, called EuroWordNet [Gilarranz, et al. 97]. These semantic taxonomies in EuroWordNet, have been developed for Dutch, Italian and Spanish and are planned to be extended to other European languages. Related activities have been launched in Europe, such as ACQUILEX⁶ (Acquisition of Lexical Knowledge for Natural Language Processing Systems), ESPRIT MULTILEX⁷ (Multi-Functional Standardised Lexicon for European Community Languages).

⁴ <http://morph ldc.upenn.edu/ldc/news/release/hansard.html>

⁵ <http://www.hd.uib.no/icame.html>

⁶ <http://www.cl.cam.ac.uk/Research/NL/acquilex>

⁷ <http://www.twente.research.ec.org/esp-syn/text/5304.html>

27 The collection may contain parallel and/or comparable corpora. A parallel corpora is a collection which may contain documents and their translations. A comparable corpora is a document collection in which documents are aligned based on the similarity between the topics which they address. Document alignment deals with documents that cover similar stories, events, etc. For instance, the newspapers are often describing political, social, economical events and other stories in different languages. Some news agencies spend a long time in translating such international articles, for example, from English to their local languages (e.g. Spanish, Arabic). This high-quality parallel corpora can be used as efficient input for evaluating cross-language techniques.

28 Sheridan and Ballerini developed an automatic thesaurus construction based on a collection of comparable multilingual documents [Sheridan, Ballerini96]. Using the information retrieval system Spider this approach has been tested on comparable news articles in German and Italian (SDA News collection) addressing same topics at the same time. Sheridan and Ballerini reported that queries in German against Italian documents achieve about 32% of the best Spider performance on Italian retrieval, using relevance feedback. Other experiments on English, French and German have been presented in [Wechsler, Schäuble98]. The document alignment in this work was based on indicators, such as proper nouns, numbers, dates, etc. There is also alignment based on term similarity as in latent semantic analysis. This allows mapping text between those documents in different languages.

Indexing by Latent Semantic Analysis

29 In previous approaches the ambiguity of terms and their dependency leads to poor results. Latent Semantic Indexing (LSI) is a new approach and a new experiment in multilingual information retrieval which allows a user to retrieve documents by concept and meaning and not only by pattern matching. If the query words have not been matched, this does not mean that no document is relevant. In contrast, there are many relevant documents which, however, do not contain the query term word by word. This is the problem of synonymy. The linguist will express for example his request differently as computer scientist. The documents do not contain all possible terms that all users will submit. Using thesauri to overcome this issue remains ineffective, since expanding query to unsuitable terms decreases the precision drastically.

30 The latent semantic indexing analysis is based on singular-value decomposition [Deerwester et al. 90]. Considering the term-document matrix terms and documents that are very close will be ordered according to their degree of "semantic" neighborhood. The result of LSI analysis is a reduced model that describes the similarity between term-term, document-document, and term-document relationship. The similarity between objects can be computed as cosine between their representing vectors.

31 The results of the LSI approach have been compared with those of a term matching method (SMART). Two standard document collections MED and CISI (1033 medical documents and 30 queries, 1460 information science abstracts and 35 queries) have been used. It has been showed that LSI yields better results than term matching.

32 Davis and Dunning [Davis, Dunning95] have applied LSI to cross-language text retrieval. Their experiments on the TREC collection achieved approximately 75 % of the average precision in comparison to the monolingual system on the same material [Davis, Ogden97]. The collection contains about 173,000 Spanish newswire articles. 25 queries have been translated from Spanish to English manually. Their results reported in TREC-5 showed that the use of only dictionary-based query expansion yields approx. 50 % of the average precision in comparison to results of the multilingual system. This degradation can be explained by the ambiguity of term translation using dictionaries.

33 This technique has been used by Oard [Oard96] as the basis for multilingual filtering experiments, and encouraging results have been achieved. The representation of documents by LSI is "economical" through eliminating redundancy. It reduces the dimensionality of document representation (or

modeling), and the polysemy as well. However, updating (adding new terms and documents) in representation matrices is time-consuming.

2.4 Related Projects

34 In the following, selected projects of multilingual text retrieval will be described.

EMIR

35 A combination of machine translation and other information retrieval methods has been applied in the EMIR project⁸ (European Multilingual Information Retrieval) [Fluhr 90] and [Fluhr et al. 97]. It aims at extending the text retrieval system (called SPIRIT, see below), developed by Fluhr et al., in multiple languages. The EMIR system allows a user to submit a query in his/her preferred language and to get documents in another language than the query language. If the language of the document is not understandable for the user, than a translation of the documents is useful. Even if the user is fluent in all languages, it is impossible to submit a query in all languages in order to get the overview of the whole collection.

36 The following components have been developed in this project:

- linguistic processors for morphological and semantic analysis for the text and the query
- a statistical model to weight word-document relationship

37 In EMIR, the linguistic process is responsible for parsing documents and queries. The terms will be lemmatized and weighted using their co-occurrence values. In case of query terms, new possible terms will be derived using monolingual and bilingual rules in order to expand the query. Other rules help recognize compound terms and translate them properly. To this end, bilingual dictionaries of compound terms and so-called multi-terms, such as “looking forward”, are very helpful.

38 The textual database is used as filter to separate correct translations from the wrong ones. Translation ambiguities are eliminated or reduced by comparing them with the textual database and with the help of co-occurrence frequencies. For instance, the word “traitement” in French has many translations in English (treatment, processing, salary, etc.). Suppose that the scope of our database deals with computers and technology, then the translation “processing” seems to be the suitable one. If the domain of the database is medical-oriented, then “treatment” should be weighted highly.

39 In this project an evaluation has been based on the English Cranfield collection with 1398 aeronautical abstracts. 225 queries have been translated from English into French manually. The EMIR system was 8% better than using the machine translation Systran followed by monolingual retrieval [Radwan94]. The outcome of the EMIR project is the commercial product SPIRIT in multiple languages.

SPIRIT (Syntactic and Probabilistic System for Indexing and Retrieving Textual Information)

40 SPIRIT has been built over a 100 man-years R&D effort carried in cooperation with the CEA (Commissariat à l’Energie atomique, French Atomic Center) and other European research organisms. The user can submit his queries in a natural language, e.g. in complete phrases or sentences. No requirements are needed to learn how to formulate queries. Boolean operators are also supported.

⁸ <http://www-uk.research.ec.org/esp-syn/text/5312.html>

41 SPIRIT used the results of the EMIR project and provides a user with multilingual search possibility. That means, a user can input his request in one language and he will get documents in other languages. Today, English, French, German and Russian are supported. The morphological analysis consists of identifying word and phrase in a text and giving its related stems and grammatical categories. To this end, SPIRIT contains about 350.000 forms in English, 500.000 in French and 1.200.000 in German. SPIRIT provides also a syntactic analysis in order to recognize the syntactic ambiguities for a given term and also to identify the syntactic dependencies between the words.

Spider (EuroSpider)

42 The Spider (later a commercialized product EuroSpider⁹), developed at the Swiss Federal Institute of Technology, is a multilingual information retrieval based on thesaurus-based query expansion approach performed over a collection of comparable multilingual documents.

43 The Eurospider retrieval system is based on fully automatic indexing (no manual indexing required). The EuroSpider system has been evaluated at TREC-5 [Harman96]. It provides many functions of the new generation retrieval systems such as relevance ranking, word normalization, relevance feedback, automatic indexing. Eurospider's architecture allows powerful integration of database management systems and advanced retrieval functions. When adding the Eurospider system to an existing database system, the database applications do not have to be changed. A database system and the Eurospider retrieval system provide complementary functions constituting a good combination.

MULINEX (Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web)

44 The MULINEX project¹⁰ aims at providing an efficient management component of multilingual online information. This project is coordinated by DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH) in Saarbrücken, Germany.

45 In the initial phase of the project search queries and summaries of web documents are translated to enable the user to make use of multilingual information. Later, computer linguistic tools will be used to index the multilingual content according to its concepts automatically. This will benefit multilingual information providers and services increasingly.

46 Features of MULINEX are described in the following:

- Search results can be filtered according to a language and subject area.
- Concept-based search avoids irrelevant search results which often arise due to ambiguities in a language (e.g. bank as either a financial institution or a park bank).
- Automatic translation facilitates the understanding of documents in foreign languages. The efficient administration of multilingual WWW sites lowers costs for the provider and increases the user acceptance.

CANAL (Catalogue with Multilingual Natural Language Access /Linguistic Server)

47 CANAL/LS¹¹ is a project supported within the framework of the LIBRARIES program of the European Union, and is coordinated by TEXTEC Software¹² in Germany. The objective of the CANAL / LS project is to allow a user to search in multilingual online catalogs. It provides the following functions:

⁹ <http://www.eurospider.ch/>

¹⁰ <http://www.dfki.de/lt/projects/mulinex>

¹¹ <http://saarland.sz-sb.de:2222/canal/can1.htm>

¹² <http://www.textec.de>

- Query in natural language
- recognition of lemmatized forms
- recognition of compound and multi words, such as Wörterbuch and air bag.
- translation of key words into other languages.

48 A so-called Linguistic Server was necessary to analyze the user query syntactically and semantically. It is also possible to translate user query into different languages. However this translation is very simple because the context of the query terms is not taken into consideration. Analyzed and eventually translated query terms are sent to the library catalog in order to match relevant entries.

TRANSLIB (Tools for Accessing Multilingual Library Catalogues)

49 Tools for supporting multilingual access to library catalogs have been developed in the framework of the European Project TRANSLIB¹³. The project is coordinated by the KNOWLEDGE S.A. company in Greece. In this project, existing tools (e.g. machine translation, grammar parser and checker) and corpora (such as terminological databases, thesauri, electronic dictionaries) have been evaluated.

50 The developed tools help a user to access OPACs in English, Greek and Spanish. Therefore, it supports three languages (English, Greek and Spanish). The search tools are based on the multilingual thesaurus EUROVOC. The TRANSLIB system has been evaluated by Greek and Spanish librarians and has shown that improvements in user interface are necessary.

EuroWordNet

51 The main goal of this project is developing concept database (also called ontology) which provides basic semantic relations between words for several European languages, like Dutch, Italian and Spanish. Existing resources and databases, like ACQUILEX and SIFT, have supported developing the EuroWordNet¹⁴ [Gilarranz et al. 97]. This reduces the construction costs and makes the database more reliable, various and feasible. Such databases will benefit multilingual information retrieval and other multilingual applications.

52 The EuroWordNet has been linked to the American Princeton WordNet for English [Miller et al. 1990]. The WordNet database consists of semantic relations between English word meanings (synsets) in which words with related meanings are grouped together. A noun is linked to all words that have a hyponymy or is-a relation or a meronymy or has-a relation with it, and a verb all words that have a hyponymy or an entailment relation with it. This project has been developed in the framework of the Telematics Applications Programme and was coordinated by the University of Amsterdam.

TITAN (Total Information Traverse AgenNt)

53 The aim of TITAN¹⁵, developed at Information and Communication Systems Laboratories in Japan, is to support Japanese people searching the Internet resources. It allows a user to submit requests to a Web search engine in English or Japanese. Statistical language recognition is used to identify the language of indexed Web pages. As results relevant documents are presented and the page titles are translated into the request language.

¹³ <http://peterpan.uc3m.es/proyectos/translib/HomePage.htm>

¹⁴ <http://www.let.uva.nl/~ewn/>

¹⁵ <http://titan.mcnet.ne.jp>

54 TITAN provides a user with more detailed information about retrieved documents, such as page title, ranking score, country of origin of that document, the format of a document (e.g. HTML, ASCII, JPEG), a document language, etc.

55 The TITAN approach demonstrated how simply cross-language retrieval has been achieved to search multilingual documents in the Web. However, the use of dictionary-based technique is not efficient enough to translate the context of user queries. Improvements, e.g. by integrating other cross-language techniques, should be analyzed. Moreover, TITAN should be expanded to other languages.

Mundial

56 Mundial¹⁶, implemented by Mark Davis at New Mexico State University, is an Internet search interface which allows searching documents in multiple languages. However, a query should be formulated in English. The user can choose in which language a query shall be translated and also to which search engine it shall be sent.

TwentyOne

57 TwentyOne¹⁷ is a EU funded project that is coordinated by the University of Twente in the Netherlands. It aims at developing a tool for efficient dissemination of multimedia information in the field of sustainable development. The project started in March 1996 and will be finished at the end of 1998. TwentyOne will present an efficient basis for the organizations to disseminate their information resources and to open a large market and user community to access them easily and without any language restrictions. The following features should be addressed in this project:

- Support for multilingual and multimedia documents
- Cross-language querying
- Partial translation of retrieved documents

58 The TwentyOne project combines a partial document translation with query translation using a machine translation approach. A database contains documents in Dutch, French and German. An extension to other European languages is planned.

3. Multilingual Text Access

59 The WWW has been established as widely used platform for information systems. Many companies or, in general, information providers tied their databases to the Web in order to make their information worldwide accessible and to benefit from the ease-use of Web browsers. However, there is often a problem presenting for example a foreign home page written in "foreign" languages and non-Western languages, such as Arabic, Greek, etc. We cannot expect that every user should install fonts for all character sets in order to display documents written in Arabic, Hebrew, Greek, etc. The well-known Web browsers support mainly ISO-8895-1 (see Section 3.1) and several western language specificity. Some local solutions have been implemented by Web browser providers to meet the national and local needs. However, these browsers are usually limited to a few languages only, e.g. Arabic Web browsers cannot display documents written in Greek.

60 Search engines will search for what a user types in, sometimes using fuzzy logic for stemming and query expansion. But how can you search for documents written in Arabic if your terminal does not support Arabic input? Think of name giving for organizations, buildings and so on. In most cases the

¹⁶ <http://crl.nmsu.edu/users/madavis/mundial.html>

¹⁷ <http://twentyone.tpd.tno.nl>; <http://twentyone.tpd.tno.nl/cgi-bin/21main/21demo/english/21main.html>

names are given in the local language. Many OPACs are accessible via the Internet, but it is difficult to read records from a foreign terminal if they are encoded in non-ASCII codes.

61 Some tools and applications are based on the Unicode which seems to resolve the character set and data exchange problems. Unicode seems to be the Santa Claus for character set and data exchange problems. However, migrating of legacy data should be loss free and at minimum of costs. Unicode is a single 16-bit which allows encoding of more than 65,000 characters. This means that most known languages in the world are covered by this code. Most operating systems, Microsoft, IBM, DEC, Sun and Apple use Unicode. HTML 4.0 supports Unicode as the reference character set for Web pages. Alis Technology produced Tango which supports all business languages for display and input purposes. Accent's Multilingual Mosaic is based on Unicode. The Microsoft Front Page Editor supports Unicode as well. Java was designed with Unicode. Database systems Oracle, Sybase, Informix, Adabas provide Unicode support.

62 The relevant aspects for the internationalization and multilingual text access are character sets, user interfaces and WWW which ensure correct data representation, interpretation, manipulation and presentation. According to these issues, many working groups have been built. For instance, at the W3C Consortium¹⁸ several groups focus on the internationalization of HTML, URL, HTTP, etc. Other US and European initiatives look at multilingual information access¹⁹ and meta data²⁰ (see also Section 5.1).

3.1 Character Set

63 The representation of the information character has been started by IBM with BCD (Binary Coded Decimal) standard of 63 characters. In 1964, this set has been expanded to EBCDIC (Extended Binary Coded Decimal Interchange Code) with a repertoire of 255 characters which include accents, umlauts, and other European and Latin American diacritics. Afterward, the ASCII code showed its efficiency and simplicity, and began to challenge EBCDIC. A conversion between both formats is not suitable, since the ASCII code do not deal with special characters such accents, umlauts etc.

64 In Europe, Teletext has been defined to develop the character code to meet all European languages requirements. However, it was very difficult to integrate Teletext in the existing computer solutions and finally the attempt failed [Sadowsky98].

65 Other attempts have also been made in the Middle East and the Far East. The result is that many standards exist, even for one language there were and still are many standards (Microsoft Standard, IBM standard, X Standards etc.). Local and ad hoc solutions (character code pages, communication protocols, etc.) have been developed. Incompatibility of character set and information exchange were not considered as serious issue. But since the WWW has been invented and widely used, the term "interoperability" becomes crucial.

66 The ISO 8859 character sets were designed in the mid-1980s by the European Computer Manufacturer's Association (ECMA) and endorsed by the International Standards Organization (ISO). ISO 8859 is a full series of 10 standardized 8-bit character sets for alphabetic writing languages. The 10 character sets²¹ are listed in the following:

- ISO 8859-1 or Latin1 (West European)
- ISO 8859-2 or Latin2 (East European)
- ISO 8859-3 or Latin3 (South European)
- ISO 8859-4 or Latin4 (North European)
- ISO 8859-5 (Cyrillic)

¹⁸ <http://www.w3c.org>

¹⁹ <http://www.cs.columbia.edu/~klavans/Activities/MIA/home.html>

²⁰ <http://purl.org/DC/groups/languages.htm>

²¹ The encoding table of each set can be found at <http://czyborra.com/charsets/iso8859.html>

- ISO 8859-6 (Arabic)
- ISO 8859-7 (Greek)
- ISO8859-8 (Hebrew)
- ISO8859-9 or Latin5 (Turkish)
- ISO8859-10 or Latin6 (Nordic)

67 The full list of character sets can be downloaded from IANA²² (Internet Assigned Numbers Authority) where approximately 190 character sets are registered.

68 Well the Web has linked all parts of the world and built a platform which we call "digital earth". It is amazing how one can get very fast information about any required topic. Unfortunately there are still barriers to overcome in order to benefit from those worldwide information resources.

69 Some countries are using the Latin character set which allows 8-bit representation. This means that only 256 characters can be represented and processed by their applications. However, 8-bit is not enough for the CJK character. Therefore an extension to 16-bit was crucial which allows a representation of more than 65,000 characters. In contrast, other countries, such as the United States and the UK, are very happy with the 7-bit ASCII character. We all remember the problem of sending or receiving e-mails which contain, for instance, German umlauts or French accents. Their display was impossible because the e-mail transport protocol supported only 7-bit messages. Therefore, ASCII has been extended to 8-bit, so that the transmission of special characters is guaranteed. This does not mean that the problem of the character has been solved. Incompatibility problems still remain unresolved regarding CJK languages, Arabic, Greek etc. Thus, a new character standard was very important to guarantee the interoperability and to foster the cooperation between all countries and cultures. Many companies, organizations, research institutes etc. came together and built a consortium to discuss this issue. As consequence, the worldwide character standard, Unicode, has been designed to cover all languages of the world.

70 The Unicode²³ is the worldwide character standard used for representing and encoding text to support the interchange, processing and presentation of written texts of multiple languages. All Unicode character codes are described by descriptor. Each descriptor is prefixed by "U+" and followed by hexadecimal number which corresponds to the character position in Unicode. For instance, U+0661 is the Unicode character 'ARABIC-INDIC DIGIT ONE' and U+0000 to U+00FF covers the range of Latin-1 character set. The Unicode Standard also includes punctuation marks, diacritics, mathematical symbols, technical symbols, arrows, dingbats, etc.

71 It is fully compatible with the International Standard ISO/IEC 10646-1, and contains all the same characters and encoding points as ISO/IEC 10646. The Unicode Standard also provides additional information about the characters and their use. About 65,000 characters can be encoded in Unicode which are sufficient for representing the most known character sets of the world including Arabic, Cyrillic, CJK, Latin, Thai, Tibetan, etc. Furthermore, the Unicode Standard and ISO 10646 can be extended to so-called UTF-16 (Unicode Transfer Format) that allows encoding a million more characters, without use of escape codes.

3.2 User Interface

72 A multilingual user interface must be able to support the character sets and encodings of any document. Rendering of multilingual text is a big issue. For instance, there is often the problem that Arabic documents written in one system (e.g. Macintosh) can be not displayed on some browsers, because of the existence of a number of different code pages for the Arabic language. That means, an Arabic glyph or letter has a different character code among those various code pages. This is one of the biggest issues of developing interoperable Arabic applications that can be used worldwide, such as

²² <ftp://ftp.isi.edu/in-notes/iana/assignments/character-sets>

²³ <http://www.unicode.org>

an Arabic Web browser. Many publishers, like news agencies, still provide their information as image via WWW, because no real software can help them to publish and to display their information resources correctly and properly. This solution leads to user frustration because of slow download and search inability.

73 WebFonts is a promising technology for browsing documents in foreign languages. The required set of fonts to view a document will be loaded from the font server automatically [Chase et al. 97]. At ULIS (University of Library and Information Science²⁴) in Japan, MHTML has been developed to allow a user displaying multilingual document in any Java-enabled Web browser without installing required fonts [Maeda et al. 98].

74 It must be able to display the information in multiple languages properly and easily. Additionally, the user should be able to submit or manipulate data. Thus, a accurate input facility in multiple languages is also a crucial feature of any multilingual application. Nowadays, there are some solutions which provide a virtual keyboards, such as Tango. Other products are based on research results in transliteration/transcription. As an example, we cite here the extension of MHTML from ULIS in Japan to input functions [Sugimoto et al. 98].

3.3 Internationalization Of WWW (HTML, URL, HTTP)

HTML (Hypertext Markup Language)

75 Internationalization of HTML means that HTML should be able to deal with non-Western characters in the text, such as Arabic, Chinese, Thai, etc. The non-Western characters should be represented in a HTML document properly. Further, it should support their display and other operations correctly, since a HTML document will contain text fragments in multiple language [Yergeau et al. 97]. The former HTML versions were based on ISO Latin 1 which support only West-European languages. Special characters for French, German, and Spanish languages (ISO-8859-1/Latin 1) are supported. Thus, rendering special characters like ç, Û or ñ can be performed easily and do not require any special tools. For instance, the French word “français” is encoded as “français” and all web browsers can render it correctly. How other ISO Latin 1 special characters are defined, can be found at the Center for the Study of Languages site²⁵.

76 Later, options have been added to support other European languages (e.g. ISO- Latin 2), but the real problem of the internationalization remains still open. Discussions have been performed by the W3C and the Unicode Consortiums. It came out that HTML 4.0 and future versions will be designed to support the Universal Standard Character Set, Unicode [Dürst]. There is an attribute to indicate the language of the whole document, paragraphs, list items, etc.

77 HTML 4.0 supports bi-directional multilingual text [Raggett et al. 98]. For instance, it is common that in the Arab world many documents contain Arabic and English or French text. Thus, bi-directional rendering support is very crucial. Because of the facilities of CSS (Cascading Style Sheets) it will be also possible to deal with bi-directional text in the new Markup language XML (eXtended Markup Language). There are many companies and organizations which provide their information in multiple languages. Thus, HTML should be able to create links between different translation and versions.

²⁴ <http://www.ulis.ac.jp>

²⁵ <http://www.ruf.rice.edu>

XML (eXtended Markup Language)

78 XML 1.0 is a relatively new markup language in the Web. It becomes easy to define document types which can be globally shared on the Web. According to user defined rules (e.g. such those of document types) XML documents can be parsed and validated. The separation between a document content, a structure and layout will benefit the interoperability of applications highly.

79 XML is based on ISO 10646/Unicode. XML requires that any XML processor accepts both UTF-8 and UTF-16 (Unicode encodings). It supports also the attribute (xml:lang) which indicates the language of the contents [Bray et al.98].

URL/URI (Uniform Resource Locator/Identifier)

80 URLs are Web resource addresses[Berners-Lee et al. 98]. They are limited to ASCII. However, an address is the facility and “guide” to find someone or something. This restriction to ASCII forces other people (Chinese for instance) to provide their addresses in ASCII, even if these addresses are for the local needs only. Furthermore, we close the door to the users who are not familiar with ASCII to profit from the Internet resources and services.

81 In the framework of the internationalization of WWW, some discussions were launched on this topic. Among diverse solutions discussed, the use of UTF-8 seems to be the preferred character encoding for URIs [Dürst]. UTF-8 is an encoding of Unicode into 8-bit characters. This encoding is suitable because on the one hand it allows representing URL/URIs in the local character code (Unicode-based), and on the other hand it is compatible to the current approach, i.e. the current URLs based on ASCII can still be used.

HTTP (Hypertext Transfer Protocol)

82 HTTP is 8-bit protocol. Special characters can be transferred properly. HTTP uses language tags within Accept-Language and Content-Language (RFC1766). The same language codes are used for HTML, XML, CSS. However, a few compatibility issues, such as MIME type (Multipurpose Internet Mail Extensions) while transmitting Unicode text, should be addressed in the future. Furthermore, the HTTP protocol ignores in some implementations the charset parameter (e.g. Charset =ISO8859-1 or charset=US-ASCII) which leads to incompatibility problems by the clients and finally to mis-handling of the information. In general, HTTP seems to be the less challenged problem of the internationalization of WWW.

3.4 Related Activities

83 This part aims at creating a web network of all known projects with links to the institutions, companies who are working on multilingual information access. Furthermore, these projects will be described according to their main research and application goals.

Tango

84 Tango, multilingual browser developed by Alis Technologies company²⁶, enables Internet users to display Web sites in more than 90 languages and provides user interface in 18 languages, like Arabic, French, German, Russian, Chinese, Japanese and Korean. Tango is based on the Universal Set Character Code Unicode and designed to run on Windows 3.x, Windows95 and WindowsNT. Some users reported that they are using it on windows98. A feature integrated within the Tango multilingual

²⁶ <http://www.alis.com>

browser is an Internet mail client that allows users to create, send and receive e-mail messages in over 90 languages.

85 With Tango 3.0, we could browse Arabic pages using code pages such as the CP 1256 of the Arabic Windows, Macintosh code pages and ISO-8859-6. The Alis staff reported to us that most of the other Arabic character sets, such as ASMO 708+, ASMO 708-FR and DOS 864 can be displayed, which we unfortunately did not test.

86 Tango allows the users to switch between the user interface languages smoothly and without new starting browser. It provides Help files in various languages, including Arabic. An integrated virtual keyboard offers input capabilities in different languages. Tango offers input capabilities that allow users to submit a query in search engines or forms in languages other than English. Other facilities allow to change the direction from left to right, and right to left, and to change the font size. There is no way to change fonts, though.

87 Tango really is, as we know, the unique product that allows Web users to show the "foreign" Web pages in most known business languages, in particular in Asian languages, Arabic and other "exotic" languages. While other products focus on specializing in certain languages pairs or triple. It is unfortunate that this product is available just for Windows and does not support Java and JavaScript.

88 **Tango Creator** is a multilingual Web editor that helps developers build Web pages in their local languages. It is based on Unicode. Tango Creator allows complex characters such as Japanese Kanji characters to be entered into the same HTML page as English text. Over 50 virtual keyboards are incorporated.

89 Some scripts can be displayed easily because a single character corresponds to a single glyph and the rendering order of such glyphs does not change. But in case of other scripts, such as Arabic, Hindi etc. sophisticated rendering tools are necessary. For instance, an Arabic character code corresponds up to the position mostly to four glyphs (at the beginning, in the middle, at the end of the word, and as a standalone character). Moreover, the sequence of rendering character differs from English.

90 The rendering tools, called Flores toolkit, and the Batam library used in Tango browser and Tango creator help to solve such problems. They are also based on the Unicode character set. Therefore, they help developing multilingual applications by supporting different scripts, bi-directional rendering, etc. Additionally, Flores and Batam provide other facilities, like mouse selection, hyphenation, character fonts, line breaking etc.

MHTML

91 This solution, developed at the University of Library and Information Science (ULIS) in Tsukuba Japan, aims at presenting multilingual documents, even if a browser and a client platform do not contain and support the required fonts. This system has been extended to allow the input in multiple languages. Using MHTML Server users can display and search multilingual documents from any Java-enabled browser [Maeda et al. 98].

92 The MHTML system²⁷ consists of two components, MHTML server and MHTML viewer applet. The MHTML server converts on the fly a HTML document into a MHTML document and sends it to a client. Once the viewer applet receives the MHTML document, it will display it on the client browser. Both components are described in more detail in the following:

93 **Multilingual-HTML (MHTML) server** is a WWW gateway between a Web client and a Web document server which converts a HTML document requested by the client into a MHTML form. The requested document in the MHTML form is displayed on the client by a Java applet, called MHTML

²⁷ <http://mhtml.ulis.ac.jp>

viewer. The MHTML document contains the source HTML text and the minimum set of font glyphs needed to display the source document. The server sends the Java applets and the MHTML document to the client. Since the MHTML viewer uses the font glyphs contained in the MHTML document, no fonts of “foreign” languages are necessary.

94 **The MHTML applet viewer** is a Java applet which allows displaying HTML pages written in various languages, without the installation of any fonts. The use of this applet allows to display multilingual documents easily on WWW.

95 Recently, the MHTML solution has been extended to an **Input function** for “foreign” texts [Sugimoto et al. 98]. A text input function is a mapping from a key input sequence to a character code or to a character code string. The mapping function can be located in a remote server as well as the MHTML server which makes a character code string in a foreign language visible to the user.

96 The text input function will provide a user with the ability to search in multilingual documents, even if their computers do not include or support corresponding input systems. The user submit his/her query in transliterated Japanese which will be converted to the Japanese characters. The converted terms are presented to the user in order to be verified (feedback). Finally, the selected terms will be sent to search engines to retrieve a collection. To avoid the drawbacks of the transliteration, a virtual keyboard to map the characters could achieve good results. This function is first tested for Japanese and it is planned to be expanded to other languages.

97 MHTML has been installed in Virginia Tech Library and has been used to view an electronic collection of folktales in multiple languages [Dartois et al. 97]. MHTML is now restricted to CJK character set, western languages, Cyrillic, Greek, Turkish and Thai. Recently, discussions have been launched by the MHTML team from ULIS Japan and the author in order to extend the multilingual browser to the Arabic language.

98 The source code²⁸ of MHTML is available for download free of charge. The following platforms are supported:

- Solaris 2.3, 2.4, 2.5.1, 2.6
- SunOS 4.1.3, 4.1.4
- HP-UX 9.07
- IRIX 5.3, 6.2, 6.3, 6.4
- OSF1 V3.2
- Linux 2.0.30
- BSD/OS 2.0.1
- NEWS-OS 4.2.1R (libmissing.a is required)
- NEWS-OS 6.1

Accent

99 Accent²⁹ is a plug-in for Netscape Navigator (2.0-3.0) which lets a user browse in documents written in over 30 languages. It runs under Windows 3.1x and Windows95. Help files and a user interface are available in few languages. However, Accent browser has not been updated since two years.

Sindbad

100 Sindbad, developed by Sakhr Software company³⁰, is an extension of Netscape Navigator that runs under the Arabic and Latin versions of Windows 95. Beside the facilities that “Latin” Netscape

²⁸ <http://mhtml.ulis.ac.jp/download>

²⁹ <http://www.accentsoft.com>

³⁰ <http://www.sakhr.com>

Navigator provides, Sindbad plug-In allows displaying Arabic documents and sending and receiving e-mails in Arabic. However, the e-mail facility and other Sindbad features (Web publishing, chatting, conferencing, etc.) are not provided with the free version of Sindbad. Therefore, we could not test them and we could not supply any judgements here.

101 Sindbad supports the following code pages ASMO 449, 708 and those from Microsoft, Macintosh, IBM, Al-Mousaid, Al-Arabi, Nafitha, and Sakhr. In the case of garbled text, the user can easily switch between these code pages in order to find the right coding mechanism. However, displaying Arabic text with vowels (short vowels) remains the challenge of this browser and other Arabic computing systems.

102 To benefit from this solution, a user has to download the free version of Sindbad 3.0 (see below the Web site) and Netscape Navigator 3.0 (Latin version), if the latter is not already installed. Installation instructions and information about previous or newer releases can be read at the download site (see below). With this solution, Sindbad Netscape 3.0, a user can browse the Web as he does with “normal” browsers. The additional benefit is that he can browse Arabic Web sites and documents and he does not need a separate browser. The browser is available for download free of charge from the Sakhr site³¹.

IBabble: A Synoptic Unicode Browser

103 IBabble³², under development at the Institute for Advanced Technology in the Humanities, is an SGML-enabled synoptic browser that allows displaying multilingual texts, using mixed character sets. It is based on Unicode. IBabble can be run as a helper application to Netscape or other Web browsers. It is also planned to provide IBabble as Java applet, once the JDK (Java Developer Kit) will be integrated in Web browsers.

104 In order to run IBabble, the Java Development Kit Version 1.1.4 or higher is required which is, as we know, free of charge for educational and research purposes. To follow the steps of IBabble development, it may be worth to have a look at its web site (see below).

Arabic Mosaic

105 AraMosaic is an extension to NCSA Mosaic Web browser in order to support beside Latin languages the Arabic script. Actually, AraMosaic can only display Arabic text which has been encoded in ISO 8859-6.

106 AraMosaic, developed by LangBox International company, is available for Unix/X11 platforms. However, implementations in other platforms, such as Windows and Macintosh, are planned. AraMosaic can be downloaded for free at the company site³³. Binary versions already exist for the following operating systems:

- SGI 5.2/5.3/6.2
- Sun Solaris 2.4/2.5, X86 2.6
- SunOS 4.1.3/X11/OpenWindows
- Linux 2.x.x/Motif 2, 2x.x/ without Motif
- DEC Alpha OSF1 3.2

³¹ <http://www.sakhrsoft.com/lproducts/linter/free.htm>

³² <http://jefferson.village.virginia.edu/IBabble>

³³ <http://www.langbox.com/AraMosaic>; <ftp://ftp.langbox.com/pub/langbox/AraMosaic>

4. Conclusion

107 In order to make multilingual information retrieval and access more effective, the following needs are summarized:

- Presentation and visualization tools
- Multilingual indexing tools
- Multilingual collection, parallel and comparable corpora and ontologies
- Terminology extraction
- Automatic construction of transfer dictionaries
- Automatic construction of dependency relation between concepts
- Clustering and summarizing multilingual document

108 Large bilingual corpora in electronic form are definitely needed. Semantic lexicons, such as WordNet and EuroWordNet, are needed for other languages. The polysemy seems to be the big problem in multilingual retrieval. The problem becomes more serious once the size of involved topics increases. Semantic indexing could achieve here more effectiveness by reducing the ambiguity.

109 Using Controlled vocabulary in some applications is enough to achieve expected results, but nowadays the expectations of other applications and users become more pretentious. The content of the parallel corpus should be closely related to the content. The alignment process will enhance an automatic dictionary building and thesaurus generation. Another important issue will be to study the integration of dictionary and corpus-based techniques. Combining different cross-language techniques could yield satisfied results. Standardization is very crucial especially for character sets and dynamic information resources. Such standards must be comprehensive and based on well referenced models and clear definitions by involving all parties.

110 In this work, we have given a survey about information retrieval and access. Techniques, commercial and public products have been reviewed. However, we did not deal with retrieving and accessing multilingual speeches and related technical challenges. Again, because of time we just mentioned machine translation tools very briefly. We believe that we succeeded to build a Web network of well-known projects and activities related to our topic. A collection of links is given in the next Section³⁴.

5. Internet Resources

This part aims at creating a Web network of well-known³⁵ activities and resources with links to the institutions, companies who are working on multilingual information retrieval and access. Furthermore, these resources will be classified regarding to their main research and application goals. A short description will be provided to each resource citation.

5.1 Working Groups and Related Activities

Association for Computational Linguistic

The Association for Computational Linguistics is the international scientific and professional society for people working on problems involving natural language and computation. The following publications are provided: the ACL quarterly journal and COMPUTATIONAL LINGUISTICS.
<http://www.cs.columbia.edu/~acl/home.html>

Center for Spoken Language Understanding (CSLU)

This site provides access to learning experiences related to language technologies and interactive systems. It provides a gateway to the many worldwide activities in language technologies—research

³⁴ All given URLs are valid. However, we can not guarantee their validity in the future.

³⁵ To our knowledge; we welcome you to report to us non-cited activities.

efforts, language resources, conferences, links to publications, demonstrations, courses and interactive tutorials resulting from work at CSLU.

<http://cslu.cse.ogi.edu/>

ELSNET

ELSNET is the European Network in Language and Speech. The long-term technological goal which unites the participants of ELSNET is to build multilingual speech and Natural language (LN) systems with unrestricted coverage of both spoken and written language.

ELSNET is one of over a dozen Networks of Excellence established by the European Commission's ESPRIT Division for Basic Research. ELSNET encourages the development of language technology in Europe by helping coordinate progress on both the scientific and technological aspects.

<http://www.elsnet.org/>

Human Languages Page

The Human-Languages Page is a comprehensive catalog of language-related Internet resources. The over 1800 links in the HLP database have been reviewed per hand (e.g. online language lessons, translating dictionaries, native literature, translation services, software, language schools)

<http://www.june29.com/HLP/>

INFO 2000 - OII (Information on the European Commission's Open Information Interchange service) INFO 2000- OII Standards and Specifications' section provides information on character sets that can be used for data interchange. It contains details of standard code for Information Interchange.

The objective of OII services is to provide standards and specification developers, product and service providers, and end-users of these products and services with an overview of existing and emerging standards and industry specifications designed to facilitate the exchange of information in electronic form.

<http://www2.echo.lu/oii>

Language Engineering Sector of the Telematics Applications Programme

The aim of Language Engineering (LE) is to facilitate working with different European languages with the use of telematics systems. Research and Technology Development work focuses on pilot projects that integrate language technologies into information and communication applications and services.

<http://www2.echo.lu/langeng/en/lehome.html>

Salt

United Kingdom Speech and Language Technology Club:

This Web site provides news and information about a worldwide speech and language technology in general.

<http://salt.essex.ac.uk/salt/>

Description	URL
Connect Magazine: Multilingualism and the WWW.	http://www.nyu.edu/acf/pubs/connect/summer96/MultiLingSum96
Global Internet Statistics (by Language)	http://www.euromktg.com/globstats/
Institut National des langues orientales	http://www.inalco.fr
International Standards Organisation	http://www.iso.ch/ISO
LINGUIST List	http://linguist.emich.edu
Multi-Lingual Information Retrieval	http://ai.bpa.arizona.edu/mlir
Multilingual Information Society Programme (MLIS)	http://www2.echo.lu/mlis
Reader Arabic Computing Mailing List	http://leb.net/archives/reader

Survey of the State of the Art in Human Language Technology Editorial Board	http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html
TREVI Project	http://www.itaca.it/itaca/en/sito_en.htm
Web Languages Hit Parade	http://www.isoc.org:8080/palmares.en.html

Special Interest Groups

Description	URL
Digital Library Forum	http://dlforum.external.forth.gr
EU and US Working Group on Multilingual Information Access	http://www.cs.columbia.edu/~klavans/Activities/MIA/home.html
SIGDAT - Linguistic data and corpus-based approaches to NLP	http://www.cis.upenn.edu/~yarowsky/sigdat.html
SIGGEN - Natural Language Generation	http://www.cs.bgu.ac.il/siggen
SIGLEX – Lexicon	http://www.cis.upenn.edu/~mpalmer/siglex2.html
SIGMEDIA - Multimedia Language Processing	http://www.dfki.uni-sb.de/~andre/sigmedia/index.html
SIGMOL - Mathematics of Language	http://www.cis.upenn.edu/~ircs/mol/mol.html
SIGNLL - Natural Language Learning	http://www.aclweb.org/signll
SIGPARSE - Natural Language Parsing	http://www.seti.cs.utwente.nl/Docs/parlevink/sigparse
SIGPHON – Computational Phonology	http://www.cogsci.ed.ac.uk/sigphon
Working Group on Dublin Core in Multiple Languages	http://purl.org/DC/groups/languages.htm

People Working in Multilingual Information Retrieval and Access³⁶

Thomas Baker	http://www.cs.ait.ac.th/~tbaker
Bruce Croft	http://cobar.cs.umass.edu/info/people/staff/croft.html
Mark Davis	http://crl.nmsu.edu/users/madavis/
Bonnie Dorr	http://www.umiacs.umd.edu/~bonnie/
Christian Fluhr	Commissariat à l'Energie Atomique, France
Hachim Haddouti	http://www.forwiss.tu-muenchen.de/~haddouti
David Hull	http://www.xrce.xerox.com/people/hull/hull.html
Judith Klavans	http://www.cs.columbia.edu/~klavans
Doug Oard	http://www.glue.umd.edu/~oard
Carol Peters	IEI-CNR (Istituto di Elaborazione della Informazione, Consiglio Nazionale delle Ricerche) Pisa, Italy
Peter Schäuble	http://www-ir.inf.ethz.ch/Public-Web-Pages/Schauble/Schauble.html
Paraic Sheridan	Swiss Federal Institute of Technology, Switzerland
Dagobert Soergel	http://www.clis.umd.edu/faculty/soergel/soergel.html
Shigeo Sugimoto	http://eboshi.ulis.ac.jp/~sugimoto/
François Yergeau	http://www.alis.com:8085/~yergeau/

³⁶ We apologize if we forget your name, please help us to complete this collection.

5.2 Tools and Demos

Description	URL
Accent Software - language solutions for the Internet	http://www.accentsoft.com/ara/accara.htm
Alis Translation Solution	http://www.alis.com/castil/flores/rendu.html
Arabic Archive (software, application etc.)	http://www-ceg.ceg.uiuc.edu/~haggag/arabic.html
Arabist: search engine	http://www.arabist.com
Arabization of a User Interface	http://www.langbox.com/staff/arastub.html
ArabTeX Version 3	ftp://ftp.informatik.uni-stuttgart.de/pub/arabtex/arabtex.htm
Club des utilisateurs Doris/Loris	http://enssibhp.enssib.fr/club-doris
Electronic Text Center: Localization of HTML	http://www.rice.edu/Fondren/ETC/local.html
Freely Available Information Filtering Systems	http://www.ee.umd.edu/medlab/filter/software.html
IMMD8 Computational Linguistics Software Archives	http://fau181.informatik.uni-erlangen.de/IMMD8/Services/lt/index.html
Konouz: Arabic search engine	http://www.konouz.com
Language Translation Software by Language Force	http://www.languageforce.com/default.htm
Mozilla Bi-Di Support - AraMosaic	http://www.langbox.com/AraMosaic/mozilla
Multilingual PC Directory for Internet font sites	http://www.knowledge.co.uk/xxx/mpcdir/inetsite.htm
Nisus Software, Arabic processing for Mac	http://www.nisus-soft.com
PC Arabic Computing Resources	http://www.rdrop.com/~abdu/arabic.html
UPLIFT Linguistic Tools for IR	http://www.wots.let.ruu.nl/~uplift
Xerox Arabic Home Page - Arabic Morphological Analysis and Generation	http://www.xrce.xerox.com/research/mltt/arabic
Yamada Font Archive, University of Oregon	http://www.gy.com

5.3 Corpora, Lexicons, Dictionaries etc.

Description	URL
Tyler Chambers, Internet Dictionary Project	http://www.june29.com/IDP
Alternative French	http://www.notam.uio.no/%7Ehcholm/altlang/ht/French.html

Arabic-English Dictionary	http://www.sakhr.com/qamoos.htm
British National Corpus	http://info.ox.ac.uk/bnc/corpora.html
Brown Corpus	http://www.hd.uib.no/icame.htm
Canadian Hansard	http://morph ldc.upenn.edu/ldc/news/release/hansard.html
CEDICT (Chinese-English Dictionary)	http://www.mindspring.com/~paul_denisowski/cedict.html
CYC (Common-sense Knowledge Base)	http://www.cyc.com/cyc-2-1/cover.html
EDICT Project	http://www.dgs.monash.edu.au/~jwb/japanese.html#edict_proj
Euralex-Dictionary and Corpora Resources	http://www2.ims.uni-stuttgart.de/euralex
European Corpus Initiative	http://www.hcrc.ed.ac.uk/Site/ECI.html
Eurovoc	http://www.psp.cz/cgi-bin/eng/eurovoc/
EuroWordNet	http://www.let.uva.nl/~ewn/
French-English	http://sun-recomgen.univ-rennes1.fr/FR-Eng.html
GermaNet	http://www.sfs.nphil.uni-tuebingen.de/lzd/english.html
Handwörterbuch Online	http://www.gsmuc.de/look.html
Information on thesaurus maintenance programs	http://www.fbi.fh-koeln.de/labor/bir/thesauri_new/thsoften.htm
Langenscheidt's T1 Translation Quality	http://www.gsmuc.de/english/trans.html
LEO-collection of Web-Dictionaries	http://dict.leo.org/dict/
LINGUIST List: Dictionaries	http://www.emich.edu/~linguist/dictionaries.html
Logos Dictionary	http://polyglot.lss.wisc.edu/lss/lang/langlink.html
LSD (German version of WordNet)	http://www.sfs.nphil.uni-tuebingen.de/lzd/english.html
MLIS Programme: The Language Engineering Directory	http://www2.echo.lu/mlis/en/leddesc.htm#What
Multilanguage Search page	http://www.logos.it/query/query.html
Multilingual thesaurus Management MTM at TermNet	http://www.infoterm.or.at/termnet.html
On-line Dictionaries	http://www.bucknell.edu/~rbeard/diction.html
Parallel Corpora	http://info.ox.ac.uk/bnc/corpora.html
SENSUS (Another variant of WordNet linked to lexicons of Japanese, Arabic, and English)	http://mozart.isi.edu:8003/sensus
Travlang front page	http://www.travlang.com/
Universal Decimal Classification	http://www.chem.ualberta.ca/~plambeck/udc/index.htm
Webster Dictionary	http://www.m-w.com/netdict.htm
WordNet	http://www.cogsci.princeton.edu/~wn/w3wn.html

5.4 Selected publications and Related resources

Description	URL
ArabTex Stuttgart	http://www.informatik.uni-stuttgart.de/ifi/bs/publikationen.html
Arab Information Project - Georgetown University	http://www.georgetown.edu/research/arabtech
Babel: Internationalization of the Internet	http://babel.alis.com:8080
Code table of ISO8859 series	http://czyborra.com/charsets/iso8859.html
Colibri: An electronic newsletter on language, speech and logic	http://colibri.let.ruu.nl
Connect from ACF	http://www.nyu.edu/acf/pubs/connect/
Cross-Language Information Retrieval Resources	http://www.ee.umd.edu/medlab/mlir/mlir.html
Cross-Language Information Retrieval Resources	http://www.clis.umd.edu/dlrg/clir
Electronic Text Center: Localization of HTML	http://www.rice.edu/Fondren/ETC/local.html
EMIR	http://www.newcastle.research.ec.org/esp-syn/text/5312.html
GIPSY: Automated Geographic Indexing of Text Documents	http://bliss.berkeley.edu/papers/gipsy/gipsy.html
IICA: An Ontology-based Internet Navigation System	http://ai-www.aist-nara.ac.jp/doc/people/mitiak-i/aaai96/
Internationalisation of HTML	Http://www.alis.com:8085/ietf/html
Internet surveys	Http://www.nua.net/surveys/how_many_online/index.html
ISO 639 Languages and Dialects	http://www.colba.net/~mgelinas/iso/lang-en.html
Language, Alphabets, and the Multilingual Internet	http://www.nyu.edu/acf/pubs/connect/summer98/DirMultiSum98.html
MULINEX project summary	http://www2.echo.lu/langeng/en/le3/mulinex/mulinex.html
Multilingual Information Processing	http://www.etl.go.jp/~mule/symposium/welcome.html
Multilingual WWW	http://mirage.irdu.nus.sg/multilingual/unicode/misc/multilingual-www.html#ID-97DB1FD1
Nota Bene	http://www.notabene.com
Nua Internet - How Many Online	http://www.nua.net/surveys/how_many_online/index.html
Survey of the State of the Art in Human Language Technology	http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html
Text categorization for multiple users based on semantic features from a machine-readable dictionary	http://www.acm.org/pubs/toc/Abstracts/tois/183425.html
Unrefereed Publications	http://www.cs.duke.edu/~mlittman/docs/unrefer.html
Working notes of the AAAI Symposium on Cross Manguage Text and Speech Retrieval	http://www.ee.umd.edu/medlab/filter/sss/papers

6. References

- [Ballesteros, Croft] L. Ballesteros, W.B. Croft . Dictionary-based methods for cross- lingual information retrieval, Proc. of the 7th Int. DEXA Conference on Database and Expert Systems Applications, 1996.
- [Berners-Lee et al. 98] T. Berners-Lee, R. Fielding, L. Masinter, Uniform Resource Identifiers (URI): Generic Syntax, work in progress, June 1998 (<ftp://ftp.ietf.org/internet-drafts/draft-fielding-uri-syntax-03.txt>)
- [Bray et al.98] T. Bray, J. Paoli, C. M. Sperberg-McQueen, Eds., Extensible Markup Language (XML) 1.0, W3C Recommendation 10-February-1998 (<http://www.w3.org/TR/REC-xml>)
- [Chase et al. 97] B. Chase et al. Web Fonts, W3C Working Report Draft, July 1997 (<http://www.w3.org/TR/WD-font-970721> works in progress)
- [Croft95] W. B. Croft. What Do People Want from Information Retrieval? (The Top 10 Research Issues for Companies that Use and Sell IR Systems). D-Lib Magazine, November 1995 (<http://www.dlib.org/dlib/november95/11croft.html>)
- [Davis, Dunning95] M. W. Davis and Ted E. Dunning. A TREC evaluation of query translation methods for multi-lingual text retrieval. In D. K. Harman, TREC-4, NIST, November 1995
- [Davis, Ogden 97] M. W. Davis, W. C. Ogden . Implementing cross-language text retrieval system for large-scale text collection on the world wide web. In AAAI Symposium on Cross-language Text and Speech Retrieval. American for Artificial Intelligence, 1997 (<http://www.wv.ee.umd.edu/medlab/filter/sss/papers/davis.ps>)
- [Dartois et al.] M. Dartois et al. A multilingual Electronic Collection of Folk Tales Casual Users Using off-the-shelf Browser. D-lib Magazine, October 1997 (<http://www.dlib.org/dlib/october97/sugimoto/10sugimoto.html>)
- [Deerwester et al. 90] S. Deerwester, S. T. Dumais, R. Harshman. Indexing by Latent Semantic analysis. Journal of the American Society for Information Science, vol. 41, 6, Sept. 1990 (<http://superbook.bellcore.com/~std/papers/JASIS90.ps>)
- [Dürst] M. J. Dürst. The Next Topics for WWW Internationalization (<http://www.w3.org/International/martin.duerst.html>)
- [Edwards 93] J. A. Edwards. Survey of Electronic Corpora and Related Resources for Language Researchers. Ed. Edwards J.A, Lampert M. D. In "Talking Data: Transcription and Coding in Discourse Research. NJ, Erlbaum, London and Hillsdale, 1993 (<http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/doc/notes/corpora.txt>)
- [Fellbaum98] C. Fellbaum. WordNet (eds.): An Electronic Lexical Database. MIT Press, 1998
- [Fluhr 90] C. Fluhr. Multilingual access to full text databases. In International A.I. symposium 90 Nagoya, November 14-16, 1990, Nagoya, Japan, pp. 107-110
- [Fluhr et al. 97] C. Fluhr, D. Schmit, F. Elkateb, P.Ortet, K. Gurtner . Multilingual Database and crosslingual Interrogation in a Real Internet Application. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA,1997
- [Gilarranz et al. 97] J. Gilarranz, J. Gonzalo, F. Verdejo. An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database, in Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, 1997
- [haddouti98] H. Haddouti. Multilinguality Issues in Digital Libraries. Proceedings of the EuroMed Net'98 Conference Nicosia, March 3-7, 1998
- [Harman96] D. Harman. Overview of the Fifth Text REtrieval Conference (TREC5). NIST, 1996
- [Hayashi97] Y. Hayashi, et al.. TITAN: A Cross-linguistic Search Engine for the WWW, in Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, 1997

- [Hull, Grefenstette96] D. A. Hull, G. Grefenstette. Querying across languages. A dictionary-based approach to multilingual information retrieval, In Proceedings of the 19th ACM SIGIR Conference, 1996
- [Maeda et al. 95] A. Maeda, T. Fujita, L. S. Choo, T. Sakaguchi, S. Sugimoto, K. Tabata. A Multilingual Browser for WWW without Preloaded Fonts. In Proceedings of ISDL95, August 1995
- [Maeda et al. 98] A. Maeda, et al. Viewing Multilingual Documents on Your Local Web Browser, Communications of the ACM, 41(4): 64-65, April 1998
- [Miller et al. 90] G. Miller et al. Five Papers on WordNet. CSL Report 43. Cognitive Science Laboratory, Princeton University, 1990 (<http://www.cogsci.princeton.edu/~wn/>)
- [Nicol95] G.T. Nicol. The Multilingual World Wide Web, Electronic Book Technologies, Tokyo, 1995
- [NUA98] NUA, Internet Consultancy and Developer, October 1998 (<http://www.nua.net>)
- [Oard96] D. W. Oard. Adaptive Vector Space Text Filtering for Monolingual and Cross-language Applications. Ph.D. Thesis, University of Maryland, College Park, 1996.
- [Oard, Dorr96] D.W. Oard and B.J. Dorr. A survey of multilingual text retrieval, Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, 1996
- [Peters, Picchi97] C. Peters, E. Picchi. Across Languages, Across Cultures: Issues in Multilinguality and Digital Libraries D-Lib Magazine, May 1997 (<http://www.dlib.org/dlib/may97/peters/05peters.html>)
- [Powel, Fox98] J. Powel, E. A. Fox. Multilingual Federated Searching Across Heterogeneous Collections. D-Lib Magazine, September 1998 (<http://www.dlib.org/dlib/september98/powell/09powell.html>)
- [Radwan 94] K. Radwan. Vers l'accès multilingue en langage naturel aux bases de données textuelles. Ph.D. : 3081, l'Université Paris-Sud d'Orsay, IEF - Institut d' Electronique Fondamentale, February 1994.
- [Raggett et al. 97] D. Raggett, A. Le Hors, I. Jacobs, Eds., HTML 4.0 Specification, W3C Recommendation 18-Dec-1997 (revised on 24-Apr-1998) (<http://www.w3.org/TR/REC-html40>)
- [Sadowsky98] G. Sadowsky. Language, Alphabets, and the Multilingual Internet. In Connect, Summer Edition 1998 (<http://www.nyu.edu/acf/pubs/connect/summer98/DirMultiSum98.html>)
- [Sheridan, Ballerini96] P. Sheridan, J. P. Ballerini. Experiments in Multilingual Retrieval Using the Spider System. In Proceeding of the 19th Annual International ACM SIGIR, 1996
- [Sheridan, Wechsler, Schäuble97] P. Sheridan, M. Wechsler, P. Schäuble. Cross-language speech retrieval. In AAAI Symposium on Cross-Language and Speech Retrieval, 1997
- [Soergel97] D. Soergel. Multilingual thesauri in cross-language text and speech retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA, 1997
- [Sugimoto, et al. 98] S. Sugimoto et al. Experimental Studies on an Applet-Based Document Viewer for Multilingual WWW Documents - Functional Extension of and lessons Learned from Multilingual HTML. In Lecture Notes in Computer science. Ed. C. Nikolau, C. Stephanidis, ECDL '98, Crete 1998
- [TRANSLIB95] TRANSLIB. Advanced Tools for Accessing Multilingual Library Catalogues. Technical Report, Deliverable D.1.4: Evaluation of Tools. Knowledge S.A., June 1995
- [UNESCO71] UNESCO. Guidelines for Establishment and Development of Multilingual Scientific and Technical Thesauri for Informational Retrieval. SC/WS/501, Paris, 1971
- [Wechsler, Schäuble98] M. Wechsler, P. Schäuble. Multilingual Information Retrieval Based on Document Alignment Techniques. In Lecture Notes in Computer Science. Ed. Ch. Nikolau, C. Stephanidis. Second European Conference on Research and Advanced Technology for Digital Libraries ECDL '98, Crete 1998
- [Yergeau et al. 97] F. Yergeau et al. Internationalization of the Hypertext Markup Language, RFC 2070, Network Working Group, January 1997 (<http://www.w3.org/International/francois.yergeau.html>)

GLOSSARY

ASCII	American Standard Code for Information Interchange, a simple character code
Character code	A numerical code (code position, code value, code number, code set value) to each character in the repertoire, e.g. 97 is the character code of "a" in ISO 10646
Character encoding	A mechanism of presenting characters, a mapping algorithm
Character Set or repertoire	Set of abstract representation of characters
Font	Collection or repertoire of glyphs
Glyphs	Are artistic representations in typographic style to describe the appearance of a character , e.g. the character A might be presented by two glyphs as A (bold) or <i>A</i> (italic).
ISO Latin 1 or ISO 8859-1	Defines a character repertoire for West European languages. It contains beside the ASCII code, some special characters, such as accents, umlauts, etc. There are also other standards from the same ISO 8859 family, such as Latin-2 for Central and East European languages, etc.
MIME	Multipurpose Internet Mail Extensions allow to specify a media type (e.g. text, image), a subtype (text/html) and an encoding (such as iso-8859-1). It helps transmitting beside plain text images, audio and video clips by Email.
Unicode	The Universal Character Set (UCS) or ISO 10646 is 16-bit code and can encode about 65,000 characters which are sufficient for representing the most known character sets of the world including Arabic, Cyrillic, CJK, Latin, Thai, Tibetan, etc.
UTF-7	UTF-7 is an encoding of Unicode into 7-bit characters. This is a transformation format specially for Internet e-mails.
UTF-8	UTF-8 is an encoding of Unicode into 8-bit characters.

Acronyms

Acronym	Description
ACL	Association for Computational Linguistic
ACQUILEX	Acquisition of Lexical Knowledge for Natural Language Processing Systems
ASMO	Arabic Standards and Measurements Organization
BCD	Binary Coded Decimal
CANAL/LS	Catalogue with Multilingual Natural Language Access /Linguistic Server
CEA	Commissariat a L'Energie atomique
CEDICT	Chinese-English Dictionary
CJK	Chinese Japanese Korean
COBUILD	Collins Birmingham University International Language Database
CP 1256	Code Page 1256
CSLU	Center for Spoken Language Understanding
CSS	Cascading Style Sheets
CYC	Common-sense Knowledge Base
DC	Dublin Core
DNS	Domain Name Services
EBCDIC	Extended Binary Coded Decimal Interchange Code
ECMA	European Computer Manufacturer's Association
EDICT	Japanese/English Dictionary
ELRA	European Language Resources Association
ELSNET	European Network in Language and Speech Network
EMIR	European Multilingual Information Retrieval
ETH	Eidgenössische Technische Hochschule (Swiss Federal Institute of Technology)
EUREKA	Europe-wide Network for Industrial R&D
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IANA	Internet Assigned Numbers Authority
IEI-CNR	Istituto di Elaborazione della Informazione. Consiglio Nazionale delle Ricerche.
IMMD8	Institut für Mathematische Maschinen und Datenverarbeitung in Germany
ISO	International Standard Organization
IST	Information Society Technologies
JDK	Java Developer Kit
LDC	Linguistic Data Consortium
LSI	Latent Semantic Indexing
MEDLIB	Mediterranean Virtual Library
MHTML	Multilingual Hypertext Markup Language
MIME	Multipurpose Internet Mail Extensions
MLIS	Multilingual Information Society Programme
MT	Machine Translation
MULINEX	Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web
MULTILEX	Multi-Functional Standardised Lexicon for European Community Languages
NCSA	National Center for Superconducting Applications
NIST	National Institute of Standards and Technology
NL	Natural Languages
NSF	National Science Foundation
OII	Open Information Interchange
OPAC	Online Public Access Catalog
SPIRIT	Syntactic and Probabilistic System for Indexing and Retrieving Textual Information
SYSTRAN	Machine translation product of SYSTRANSOFT Company
TC46/SC2	International Technical Committee 46 / Subcommittee Conversion of Written Languages
TITAN	Total Information Traverse AgeNt
TRANSLIB	Tools for Accessing Multilingual Library Catalogues

Acronym	Description
TREC	Text Retrieval Evaluation Conference
TREVI	Text Retrieval And Enrichment for Vital Information
ULIS	University of Library and Information Science In Japan
UPLIFT	Utrecht Project: Linguistic Information for Free Text Retrieval
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
UTF	Unicode Transformation Format
XML	eXtended Markup Language