CHAPTER 6

Faces of Cyberspace
Information presentation in visual images, sound and natural language either as text or speech is gradually becoming the norm. The Internet is creating a global platform where a worldwide forum is evolving that is becoming vital to economic, social and political success. However, all these developments create problems. Examples of such problems are: access to much of the information may be available only to the computer literate and those who understand English. Also, due to the vast amount of information available, it is becoming hard to identify and select what is relevant that may satisfy a certain degree of credibility.

The development of interfaces and components that help users to identify relevant information and then present it to them in the most appropriate manner according to the information content and their cultural and linguistic backgrounds is still lagging behind the rate of information growth. Language engineering is an endeavor in which language technologies are integrated and embedded into language-enabled services and products to support business in a global context and to facilitate interpersonal communication across languages. The Role of Language Engineering Efforts now are concentrating in developing tools and standards that enable the creation of websites in several languages or in at least one non-Western language. Now an Internet standard (RFC 2070) based on Unicode supports HTML documents in practically every language.

In particular, some of the features dealt with are: Markup of bi-directional text, i.e. test where left-to-right and right-to-left scripts are mixed and control of cursive joining behavior in contexts where the default behavior is not appropriate. Also, the HTTP which is the hypertext transfer protocol in use since 1990 is being internationalized. The relevant aspects are character set labeling which ensures correct document interpretation and language negotiation which is used at a site to provide documents in the user’s language of preference. Language technology will help in designing and implementing the systems needed to effectively deal with information and knowledge in a number of ways.

Speech recognition will help in interacting with a number of devices in our own native language. Also, information could be presented by generating speech. Understanding requests and browsing the vast amount of knowledge available is an essential component that can alleviate the problem of information overload and ensures that the relevant knowledge is accessed. It will also be possible to generate and present information in different languages through automated machine translation. Language engineering will be essential for supporting global business in general and electronic commerce in particular. The success of any business will depend on the quality of information about its customers, its competitors and the market in general. The needed information has to be identified, extracted and presented in natural language either as text or speech. In general, language engineering will deliver the right information at the right time and in the language of the recipient. Automated translation together with document management will improve the quality of service in a global marketplace. In general, success in globalization requires putting emphasis on localization.

Some companies interested in international markets organize their effort in a number of activities that could be summarized as follows:
1. Developers organize a product so that linguistic components can be modified easily (internationalization).

2. Translators make the product available in different target languages.

3. Editors review each version to ensure that culture-specific items are not missed.

4. Marketing division takes care of localization to adapt the product to the local market.

The Internet is dominated now by the English language. However, due to the near future possibilities of widespread use, it is becoming essential to consider its multilingual nature. Some efforts are being conducted in that respect - e.g. the Babel project which is a joint initiative between Alis Technologies and the Internet Society. They are considering the world's 20 main languages and are trying to study the actual distribution of languages on the Internet. Although the Internet penetration is not high at the moment in some regions, this is expected to change radically in the future. As an example, results of the Internet Domain Survey for Arab Countries is shown in Table 1 for January 1998. Efforts now are concentrating in developing tools and standards that enable the creation of websites in several languages or in at least one non-Western language. Now an Internet standard (RFC 2070) based on Unicode supports HTML documents in practically every language. In particular, some of the features dealt with are: Markup of bi-directional text, i.e. test where left-to-right and right-to-left scripts are mixed and control of cursive joining behavior in contexts where the default behavior is not appropriate. Also, the HTTP which is the hypertext transfer protocol in use since 1990 is being internationalized.

The relevant aspects are character set labeling which ensures correct document interpretation and language negotiation which is used at a site to provide documents in the user's language of preference.

**Multilingual Machine Translation Issues**

There are many situations in which multilingual translation is needed. Translation from many languages into a single language will be required by large information gathering and processing organizations. Translation from a single language into many languages will be required in the context of foreign trade when operation and other manuals for industrial equipment need to be translated into the language of the countries where the equipment is to be marketed.

**Table 1 : Internet Domain Survey Arab Countries, January 1998**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Hosts</th>
<th>All Hosts</th>
<th>Duplicate Names</th>
<th>Level 2 Domain</th>
<th>Level 3 Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>ku</td>
<td>4057</td>
<td>4749</td>
<td>692</td>
<td>9</td>
<td>2925</td>
</tr>
<tr>
<td>eg</td>
<td>2013</td>
<td>16930</td>
<td>14917</td>
<td>7</td>
<td>191</td>
</tr>
<tr>
<td>ae</td>
<td>1940</td>
<td>1955</td>
<td>15</td>
<td>8</td>
<td>56</td>
</tr>
<tr>
<td>lb</td>
<td>1134</td>
<td>1377</td>
<td>243</td>
<td>7</td>
<td>59</td>
</tr>
<tr>
<td>om</td>
<td>670</td>
<td>671</td>
<td>1</td>
<td>16</td>
<td>20</td>
</tr>
</tbody>
</table>
Multilingual interlingual machine translation systems translate between a number of languages. In this approach, a universal language independent representation of text known as interlingua is developed. Therefore, the translation process is reduced to two phases: the analysis phase between the source language and the interlingua, and the generation phase from the interlingua to the target language. For successful machine translation, detailed knowledge of the languages is required at many levels: lexicon, syntax, semantics and discourse. It is very difficult to provide such linguistic knowledge for an entire language. However, if we consider only language in a particular domain much of this knowledge could be obtained. The variety of language used in a given science or technology is not only much smaller than the whole language, but is also more clearly systematic in structure and meaning. Therefore linguists and computer scientists co-operate to study the properties of such specialized languages which are called sublanguages or controlled language. Each sublanguage has a distinctive grammar even though it is related to the grammar of the full standard language. Also, the theoretical problem of relating linguistic form to communicative function comes into sharper focus when individual sublanguages are examined.

Sometimes it is essential to distinguish between two translation activities. The first one is called localization (e.g. that used for computer manuals for end users) where it is important to adapt certain parts of the content and perhaps the style of presentation to a certain cultural and linguistic environment.

The second one is called diffusion translation where the objective content must be strictly rendered in another language without addition and omission. A brief outline will now be given for efforts related to Arabic/English translation. AppTek started an English-Arabic translator named TranSphere that uses Lexical Functional Grammar together with a general dictionary having 100,000 words. A number of domain-specific dictionaries have also been developed. This system could be either standalone or part of an integrated system.

Sakhr within Al-Alamiah Group [Sakhr, 1998] is also developing a translation scheme for Arabic-English and English-Arabic. It makes use of the available tools that have been developed before for morphological and syntactical analyzers, electronic dictionaries, semantic support together with a number of other development tools. It is also developing a general platform for dealing with Arabic computations. IBM Egypt is also developing Machine Aided Human Translation schemes. ALIS, Inc. offers a solution
that integrates core language-handling technology and translation products. Electronics Research Institute in Egypt, in co-operation with the European Community is developing an English-Arabic and Arabic to English and German medical text translation.

This is done in the framework of CAT2 (Computer Assisted Translation) which uses an interlingua approach. Also, some work related to ambiguity in Arabic language processing is being studied. PRINCITRAN: In this system a large-scale lexicon is to be constructed for an interlingual machine translation system for Arabic, English, Korean and Spanish. For Arabic, the starting point was the use of the Alpnet bilingual Arabic-English online dictionary. Then automatic mappings between English glosses from Alpnet into LDOCE (Longman's Dictionary of Contemporary English) codes were performed. The codes were then converted into thematic grids which were then exhaustively hand-verified.

**Electronic Commerce and Multilinguality**

Global market opportunities are now increasing rapidly and is going to use heavily the facilities introduced by the Internet. Thus, the advent of global electronic commerce will add an economic dimension to cultural and linguistic issues. In international trade, companies that adopt a multilingual, multicultural approach are expected to gain a competitive advantage over their monolingual, monocultural competitors.

Regarding business-consumer relations, the following language issues could be stated: The seller of goods or services must be able to publish information in the language and cultural convention of the customers (multilingual electronic publishing). The buyer of goods or services must be able to find, understand and compare information in his own language (multilingual information retrieval). Both buyer and seller must be able to interact naturally and effectively in a common language or across different languages. There is a possibility that from the number of 5,000 to 6,000 languages spoken in the world today, only a few hundred will survive a century later.

The pressure on languages can take different forms: economic, social, cultural, etc. Usually, people directly affected are minorities, but their languages represent the linguistic diversity that has developed over the course of human history. Some linguists argue that language endangerment is serious with great humanistic and scientific consequences. Can new developments in the information age help in preserving some of these languages and thus save language diversity. Information revolution is the global phenomenon that has an impact on all nations.

Nobody can stay out of the Information Society. That is why nowadays problems of ethical, cultural and social type has an increasing value. Only in resolving these problems the Information Society can achieve the declared goal of globalization: to promote quality of life and sustain a cohesive development. Due to the advent of the Information Society, UNESCO faces new problems demanding to define its role in new world communication processes. That is why the UNESCO strategy should embrace the wide range of activities:
• concentration on key structures of the Information Society;
• ethic problems of the Information Society and the Cyberspace;
• elaboration of information policy;
• the impact of the globalization processes;
• specific education;
• forming networks;
• fulfilling projects for the Information Society

These problems demand thorough considerations and constant attention. Textual information remains the main form of demanded information for majority of interstate or global international networks such as INTERNET. Due to the rapid success of communication technologies active use of spoken and visual information is foreseen in the nearest future.

Nowadays there are near 3000 spoken languages in the world and only 100 of them are written. Such a variety of languages under increased value of international contacts, high intensity of information flows of economic, political, scientific, technological and social types put forward as a priority task the development of human language technologies. Multilingualism is an inevitable historical attribute of the international communication process. The notion “multilingual information society” solidly occupies the place in the range of problems discussed by computer science specialists and predefined strategic directions of many international organizations and companies dealing with information distribution and processing.

Human language technologies obtained the valid political understanding and support. At the third UNESCO Congress on Informatics and Education held in Moscow, July 1996, more than 40 countries representatives confirmed their willingness to participate in the establishing multilingual telematics support for educational needs. The relevance of linguistic and cultural aspects of the Information Society in Europe has been stressed by the European Council meetings in Corfu, 24-25 June 1994, in Cannes, 26-27 June 1995 and on the G7 Conference of Ministers in Brussels, 25-26 February 1995. In May 1995 the European Commission has adopted the upcoming Fifth RTD Framework Programme which includes points devoted to developing human language technologies.

**Technological Multilingualism**

In the multilingualism problem there exists one important factor which usually slips out in discussions of this issue. The high level of technology of the modern civilization broadens the notion of multilingualism in information society including in it artificial intellectual electronic devices. This newly arisen complex communication structure forms the Cyberspace that is marked by extremely complex processes of control and self-development. In the Cyberspace the humanitarian multilingualism interacts with the technological multilingualism of communication protocols, computer interfaces,
programming languages and data protection methods. Communication multilingualism of computing devices is to some extent being solved by introducing conventional technological standards regulating data exchange and transformation processes. Different computers in networks can transmit information using standard popular protocols TCP/IP (Transmission Control Protocol / Internet Protocol), IPX/SPX (Internetwork Packet Exchange / Sequenced Packet Exchange) or SNA (System Network Architecture). Such protocols as FTP (File Transfer Protocol), HTTP (Hyper Text Transfer Protocol), address standard URL (Uniform Resource Locator), standard slice interface CGI are in active use in the INTERNET. High expressive possibilities of Web pages rely on implementations of HTML and Java languages.

For protecting information from different forms of eavesdropping cryptographic protocols SET (Secure Electronic Transfer), SSL (Secure Sockets Layer) and some others have been elaborated. The “hardware multilingualism” aggravates the problem of the humanitarian multilingualism in the Cyberspace. It is interesting to note some resemblance in humanitarian and technical multilingualism. Communication multilingualism is stormy developing and to some extent repeats the dramatic way of human multilingualism. Here also numerous languages and dialects currently appear and die, conflicts of standards, programming languages and interfaces is an usual event.

Any technological innovation can lead to the chain, or may be it will be better to say “domino”, reaction of expressive and communication means. The Cyberspace evolution is marked by constant attempts to resolve problems of humanitarian and technological multilingualism. The general common base for advances in these areas are backed by developing theory of semiotic systems, artificial intelligence, translation methods, algorithmic analyses, evolution of discourse, technologies sustaining lingual engineering.

**Multilingualism Challenge**

Historically formed present situation presupposes two general approaches for partial resolving the multinational communication problem. The first approach relies on the idea to adopt some universal unique well understanding language for invariant multilingual communication. The second approach lays on elaborating powerful technique for simulating expressive means of one language by expressive forms of another one. There are known three main variants in implementation of the first mentioned above approach.

- For some definite restricted knowledge areas to use specialized formalized languages such as Latin for medicine, language of formulas for mathematics, etc. This approach can be applied successfully only to very restricted object areas and cannot embrace cultural or social aspect of communication.

- To create one general artificial universal language which is neutral to national peculiarities of language communication. The most known language of this class is Esperanto. The less known is Sanskrit. The main drawback of this approach as in the previous case above is its impossibility to fill the communication gap in expressing national cultural potential.
To conveniently adopt the use of one of national language. There is one well-known very strong candidate to play this role. Domination in the world processes (trade, science, technology, culture, etc.) of English speaking countries led to the huge spread of the English as a communication language. There are 1 billion and 400 millions of English speaking people in the world. Only for 400 million of them this language is native. This means that for more than 1 billion of world population this language is the second or even the third one. And we are all witnesses that domination of the English as the communication tool in the Cyberspace is spreading with more and more increasing feedback link. The positive aspect of this event is extending for other nations the rich world of English speaking countries in all its appearances. The negative aspect is that such a global language invasion contradicts to natural eagerness of other nations to preserve local cultural achievements and adequate lingual means of their presentation. Even in the field of technological documentation complete relying on not native language information can affect the reliability of elaborating systems. Moreover, due to its internal evolution and rapid changing the fixed international English language can never be framed.

The second approach, which is simulating expressive means of one language by expressive forms of the other one, is traditionally reduced to the translation process. Before the computer era a bilingual or a multilingual interpreter was a man. Under computer expansion the language translation problem has obtained a new content - an automated language translation. In its full context this problem relates to the very complex problems of formalizing natural language understanding and cognitive context which are very far from the final decision. Nevertheless there are some recent evident successes in this area. The number of commercial systems for machine translation, as well as other types such as spell-checkers, multilingual mailers, dictionaries, etc. is rapidly growing.

During the last two years a few multilingual services available within Internet appeared. Multilingual Communication Corporation (USA) is offering its powerful WEB.TRANS service for global business. The two main browsers providers for WWW, companies Alis and Globalink, include the multilingual support in their products. The Tango, multilingual browser, using the Alis technologies, is a high performance Web tool that allows users to display Web pages authored in any of over 90 languages, select the languages of its interface, automatically retrieve these pages in the language version user prefers and input a text in a wide variety of languages. Alice Technologies has brought together several leading companies in the language industry to create Columbus - a suite of communication and translation tools and services designed to offer corporations the most comprehensive, convenient and cost-effective solutions on the market. There are other well-known companies offering a variety of multilingual solutions among them: ACCENT (Israel), Ajax Software Corp. (Russia), IBM, Intersol Brea, SYSTRAN (USA).

We can also concentrate on activity of BIT Software, Inc. (Russia), as well as ABBYY Company, formerly BIT Ukraine. These software companies are famous by their multilingual products like Stylus Gigant (powerful heuristic translator from five European languages to Russian and visa versa), Lingvo (large hyper-text interactive dictionary), WebTranSite (online “transparent” Internet translator), etc. There is one interesting
frequently met feature of the multilingualism paradigm. Due to different historical reasons in one country (nation, popularity) there are two or several internal natural languages sharing the same information space. Such a situation exists in many NIS countries, Ireland, Canada, India, Benelux countries and in many others. Ukraine is the bilingual country where the Ukrainian language has the state status. In this case specific domestic features could complicate the international multilingualism problem. Furthermore, problems become really global, e.g., after 7 years of official independence, Ukraine doesn’t even have fully defined symbol code page (380), because four or five existing standards are already widespread and it’s difficult to change any of them right now. Sometimes (as in Ireland) politics can influence the multilingualism decisions. Much attention is paid to the processes of creating computer supported information infrastructure in Ukraine.

The State program of informatization in Ukraine has been approved by the government. This program is backed by some legislative basis. There are three juridical laws concerning the informatization program. The first law defines the concept of the informatization in Ukraine. The second one defines objectives of the program. And the third law determines the content of the program. We would like to stress that the our Center is the leading organization in forming the concept and the governmental program of informatization in Ukraine.

International Research and Training Center of Information Technologies and Systems. In the organization we present here intensive research and technological works in language engineering are carried out. Let us mention briefly some of them.

- **Multilingual speech technology**: The technology of natural speech recognition and synthesis in a few foreign languages is developing by the specialists of our center for more than twenty years. In 1986-1991 in frames of UNESCO contracts the multilingual (7 languages) speech dialogue system RECH-121 for CD6-ISIS Micro has been created. The works in the direction to create oral vocabularies-translators, dictating machines, machines simulating a human voice, individual speech passports, systems of multilingual speech interpreters are continued at present.

All these elaborations are based on the modeling of human intellectual activity and on the constructive analysis of digital information by means of its synthesis. The method that is used is recognized around the world.

- **Algorithmic analysis of the deep language structures**: Attempts to create systems of artificial intelligence demand the deep analyses of historical roots and evolution of language structures and human cognition processes. Such analysis has been done using new algorithmic methods. The kernel recursive structure of language sentences and literature constructions have been revealed. The algorithmic model of language constructions invariant to the natural languages has been suggested. The practical results of this research include applications to the authorship problem, psycholinguistics, associative expert systems, analyzes of myths, multilingual translators and some others.

- Automatic forming and translation of keywords in querying systems. In the navigating the multilingual Internet the problem of adequate translation of
keywords is of great importance. The direct translation of keywords from a language A to a language B is connected with high ambiguity and can complicate the search of necessary documents. More correct idea to the keywords translation is based on the comparison of texts in the language A, which correspond to the given set of keywords, with texts in the language B which could be found by means of all variants of translations of the given key-word set. In the process of looking for similarity of keywords an neural network model could be used.

The neural network methodology for implementation of this method has been elaborated. The methodology was checked in the automatic detection of user’s interests.

These principles formed the base for creating neural networks supporting processes of making decisions and pattern recognition. Particularly, works in handwritten symbols neural recognition have been started with the Japanese company WACOM and at present gives good results. The same results could be used in the multilingual Internet.

- **Multilingual Internet and Educational Problems:** Multilingual Internet is thus becoming a real need for the nearest future, and important and urgent efforts should be made to solve the global problem of bi-directional communication barriers on different levels, addressing different user groups, etc. At the same time, due to low experience of the new users in the language and telematics, additional measures should be undertaken to provide training and educational facilities for these users.

Ukraine (and some CIS countries) can be considered as a typical “country in transition” concerning production, dissemination and use of multilingual information on the global information highways. There is an urgent need to develop an effective infrastructure in the field and to adjust to international standards. On the other hand, it is ready for international co-operative projects, having know-how and qualified specialists in many fields, including computer science, linguistics, programming, education, etc. Currently the following directions are considered as a first priorities in the National Informatization Programme: Basics in Telematics (Internet), networking of information and educational resources and multilingual computer-based support in science, education and communication, creation of the distance education centre.

Our preliminary analysis shows that the CIS countries have serious needs in the field of multilingual telematics support. Russian and French-speaking users are very conscious that they waste too much time and loose essential context information when they read/write texts in another language, even if they are competent in it. To keep one language such as English as unique language of communication is not cost-effective. There is a strong desire to stay within own native languages, which does not of course except trying to learn a few others for personal communicational and cultural purposes. The following specific multilingual user’s needs are:

- support of communication at the basic level (reading and writing) via e-mail;
- opportunities to use the adequate e-mail packages to remove essential obstacles in the way of communication both in countries, and in international communication;
- analysis of the multilingual products and services in Internet;
• analysis of the machine translators (on-line and off-line), dictionaries (on-line, off-line), thesauruses, multilingual mailers, tools for multilingual WWW-pages design, etc;

• training in multilingual Internet services (free distant courses via e-mail, WWW-courses etc.);

• updated information about new tendencies and products in multilingual telematics support;

• resolve the problems of standardization for multilingual software for every site in CIS countries;

• organization of the multilingual support of users' sites and consultation services.

The problems related to supporting user and/or suppliers groups and standardization becoming more and more actual both for East users and West suppliers. To make all relevant information available to its potential users and to provide an information exchange between Internet-users from various professional fields and from different countries, relevant national sites must be equipped by tools which form friendly and comfortable environment, making communication attractive for users world-wide, developing their information needs, widening their horizons and motivating their creativity. To facilitate a world-wide exchange via Internet specific needs of the users from wide range of cultural settings should be analyzed and generalized to find an effective solution. The general goals of the multilingual telematic support for education and training are the following:

1. to provide basic language tools for effective information exchange between Internet users from all possible professional fields and different English/French and Russian speaking countries;

2. to improve accessibility of the Internet-based information for non-English speaking users;

3. to facilitate utilization of educational and training materials available in English for the French- and Russian-speaking users as well as to ensure share of such materials available on the net of the French and Russian-speaking countries;

4. to ensure exchange of educational and training information within Russian-speaking users groups directly, without necessity to pass through English.

To achieve these goals, the interested in multilingual applications sites must be equipped with tools which form user-friendly and comfortable multilingual environment. Communication within Network must be made attractive for users world-wide, developing their information needs, widening their horizons and motivating their creativity and inter-communications. The necessity of linguistic support for multilingual communication, information search and acquisition requires the use of modern linguistic tools such as spelling and grammar checkers, dictionaries, thesauri, vocabularies, elements of automatic translation and, eventually, multimedia voice encoders and decoders. To achieve the main goals for CIS countries, the translation tools for English into Russian and Ukrainian will be used.
Virtual University as the examples of the multilingual telematics support for education and training. The life-long education of the 21st century will be global from two perspectives: from a horizontal perspective, cultures are meeting each other on the surface of the globe, entering into creative encounters that will produce a international collective virtual educational environment; from a vertical perspective, they must maximise using the world knowledge's for organisation of international virtual educational environment in order to provide a stable and secure base for future development. on the national and international levels.

The general goals of the multilingual telematic support for education and training are the following:

1. to provide basic language tools for effective information exchange between Internet users from all possible professional fields and different English/French and Russian speaking countries;

2. to improve accessibility of the Internet-based information for non-English speaking users;

3. to facilitate utilization of educational and training materials available in English for the French- and Russian-speaking users as well as to ensure share of such materials available on the net of the French and Russian-speaking countries;

4. to ensure exchange of educational and training information within Russian-speaking users groups directly, without necessity to pass through English.

These problems must be solved at local, regional, national and international levels. The global aims of the IVU "Ukraine" are the following:

1. to provide inter-cultural courses, accessible to students all over the world through distance learning methods.

2. to create multilingual learning environment for telematic-based education and training

3. to stimulate learning to creative candidates in order to form complete human beings with erudition.

The multilingual telematics support for education and training should be based on innovative ways for flexible integration of modern technologies into the new solutions directed to the development of multimedia learning environment, which in its turn will contribute to the introduction of innovative solutions and skills for telematics-based training in world society, promoting new opportunities for life-long distance education. To facilitate a world-wide exchange within the Network its local/national sites should analyse specific needs of its users to find an effective solution within the Network.

One of this specific needs is the necessity of language support for multi-lingual communication and exchange of materials. To facilitate the use of samples and examples from different sources the Network should support a communication between teachers, as well as material translation into native language. This language service within local sites would not only promote the use of English-based materials outside English-speaking countries but also open the gates to teachers experience in other cultures and
languages. The use of current machine translation tools, spell- and grammar-checkers, vocabularies would help to improve them making a necessary office tool, as text editor or word processor. Multilingual Internet support development. Main research direction is methods development for multilingual Internet technologies teaching to insure intensive use of the modern information resources by the different groups of people.

Working in Internet multilingual environment create a large amount of the problems for non-English speakers. Such users cannot to solve these problems always self-sufficiently. Activity in this direction was initiated at Second International Congress UNESCO and was supported by the 42 countries round the world. Such developments are absent in Ukraine.

However, there is strong need to overcome language barriers and ensure certain translation or explanation of terms are selected by the user. Besides that, there isn’t any materials about language barriers overcoming, in spite of existing such experience for English, Dutch and French languages. For Russian and Ukrainian languages we have only American translation off-line centers. So this problem is actual for Ukraine, that only tries to occupy certain place among leading countries in telecommunication issues. Object of investigation is ways of access to modern information Internet resources with on-line and off-line electronic dictionaries and translators, and insuring flexible distance learning and providing measures, that will facilitate active use of the multilingual Internet environment.

Multilingualism: A Canadian Experience

Since its birth a generation ago, the Internet has been dominated by the English language and North American culture. In a 1998 survey conducted by the Internet research group, eMarketer, two-thirds (68 percent) of a little over 60 million Internet users worldwide reside in just two countries, the United States, with 37 million users, and Canada, which has just over 4 million. About 60 percent of the Internet host computers are located in the United States. Nine out of 10 Internet users today are English-speaking. No fewer than 82 percent of home pages (web sites) are in English, according to the Internet Society’s survey of 60,000 computers with Internet addresses. Yet some foresee an end to this electronic hegemony.

The number of non-U.S. Internet users is about to outnumber those inside the country soon and increase by nine-fold over the next five years, from 16.4 million in 1997 to 143 million by the year 2002, representing an annual growth rate of 70 percent. In that case, the present practice of conducting business, presenting news and information, and performing discussion on the Internet will have to be drastically changed. The widespread use of English will eventually be contested and the Internet itself will become multicultural. This is already happening. A consortium of American computer companies has developed a universal digital code known as Unicode to allow computers to represent the letters and characters of virtually all the world’s languages. Major search engines like Yahoo and Excite offer their services in multiple languages.
Netscape Communications in partnership with the leading Latin American Internet service, Star Media Network, provides a free Internet guide in Spanish and Portuguese. Internet services in languages other than English, like Star Media, are starting to provide world and regional news, weather, stock listings, e-mail, chat rooms, Internet access and more, all in the users' native language. Given such developments, optimists argue that far from ending diversity, the Internet will promote it by allowing even small groups of people to disseminate their messages worldwide. By overtaking the "middle range" languages, it may actually protect minority languages threatened with extinction. A wider range of languages on the Internet means at least in theory that a wider range of ideas will be exchanged in a cyberspace, the long-promised global village. Despite a tremendous influx of non-English languages in recent years, however, the Internet has a long way to go before it becomes a truly multilingual medium.

As long as English can be understood by the largest number of Internet users, the cyberspace will continue to be dominated by English as the primary language for international discourse and commerce, European languages as a tool for regional and specialized communication, and many other minor languages for local communication. At a glance there is an advantage to having one language dominant in the cyberspace. In a world of five to eight thousand different ethnic groups who reside in approximately 160 nation states speaking 5,000 distinct languages, some language must be the common language of the Internet.

Many people believe that English must be by default the standard language on the Internet, much as it is the international language of aviation. At present there are clear indications of English becoming a lingua franca in the cyberspace. The Internet, the network of networks, is most likely to be governed by the logic of the "Metcalf's Law," the so-called magic of interconnections. Connect any number of machines, 'n', whether computers, phones or even cars, and you get 'n' squared in exponential value.

Simply put, the values of the interconnected entities increase exponentially. To expand this logic further, as English-speaking countries have happened to take the central role in world politics and economy in modern period (namely, Pax Britannica and Pax Americana), the English language will deepen its hold on the world as more people go on line. As a recent British Council report shows, the evolution of the English language accelerates as it spreads beyond Anglo-Saxonism.

The critical mass of English amongst providers and users is likely to drive the usage inexorably towards monoglot English. Even the French, famously fastidious about guarding their language against dilution by foreign words and phrases, has been forced to surrender to American vernacular on Internet matters. As English continues to dominate the cyberspace with greater intensity and speed, global resistance to it accelerates with equal strength. Some countries already disgruntled by the encroachment of the American culture—from pop music, blue jeans, to videos—are worried that their cultures will be further eroded by an American dominance in the cyberspace. In their minds, English, by association, immediately evokes a negative image of colonial imperialism. They feel that it may threaten their cultures, their languages, even identities.
For this ideological reason, many people in the Third World countries are opposed to having English as the lingua franca for commerce and trade, crisis management, and scholarly and intellectual discourse on the Internet. While the Internet is awash in information, almost all of it flows in one direction, i.e., from the United States to the rest of the increasingly wired world. It brings home the age-old controversy regarding unbalanced and unidirectional flow of information from the world's richest countries to the Third World, an on-going controversy that prompted UNESCO to declare a New World Information Order (NWIO) on several occasions.

For the ideological as well as practical purposes, many countries have taken steps to protect the Internet from excessive American influence. The French government once filed a lawsuit against a Georgia Tech campus in Metz, France, whose web page did not translate its contents into French. Arab countries proposed creation of an Arab Intranet, a closed network in which indiscriminate access to pornography and political discussion could be blocked. Canadian content requirements are to be extended to the Internet to keep out the American “cultural vulture.”

**Multilingualism: A Korean Experience**

American culture has continued to exert its influence in Korea, first through Christian missions, trade and Hollywood productions, and more recently via the Internet. Thanks to the Korean government’s aggressive drive, an estimated 2.5 million Koreans are linked to the global medium. An equally large share of the credit for the phenomenal growth of the Internet goes to the media organizations in Korea, which have in recent years vigorously campaigned for introducing the “Net” to the school community through such programs as the KidNet (Chosun Ilbo) and Internet in Education (Joongang Ilbo).

In the vicissitudes of hopes and fears about the Internet, two isolated but interrelated events recently took place in Korea, all with a nationalistic overtone. One is the Microsoft's abortive attempt to buy out a local Korean software company in exchange for a multi-million dollar investment. The other has to do with a somewhat radical proposal by a Korean writer to adopt English as a common language on the Internet.

In the midst of currency crisis in Korea these days came the shocking announcement that the Hangul and Computer (H&C) Company, a Korean software company, had agreed with Microsoft to stop development of its popular Aera A Hangul word processor program in exchange for a $20 million capital investment. Apparently, what Microsoft hoped for in the deal was to secure a monopolistic right of its Korean language version of Windows and Word in the Korean market. When it first appeared in 1990, Aera A Hangul--named after the Korean alphabet invented by King Sejong more than five centuries ago--was the only word processor based on hangul instead of Roman letters with the hangul mapping.
Arae A Hangul, "insanely great" in the words of Apple Computer's founder Steven Jobs, was multilingual at a time when DOS versions of Word and WordPerfect barely went beyond ASCII. This made it easy to develop a number of attractive fonts frequently used in displaying old Korean characters.

Media reports of the deal aroused anger and frustration among Korean computer users and non-users alike. Defenders of Hangul, who were opposed to the Microsoft's planned equity participation in H&C, quickly invoked the issue of Korean national pride to rally the public support for the program. A massive nationwide campaign saved Hangul at the eleventh hour from the threat of extinction as its developer, H&C, reversed its earlier decision to sell off the software, bowing to a nationwide campaign spearheaded by Korean scholarly community and venture capitalists. As hindsight, invoking Korean pride as a defense against predatory foreign companies appears to be too far-fetched and even xenophobic. On the other hand, Microsoft's aggressive marketing strategy to effectively terminate H&C's further development of Hangul can be viewed as too selfish and insensitive to the Korean nationalistic sentiment. Now that 2.5 millions of Koreans are linked up to form a global community, how will they communicate? Majority of Korean Internet users was found to use the global medium mainly for communication within the national boundary, according to an Internet user survey. To most of them the Internet remains largely an unexplored reservoir of knowledge and information due to language barrier.

Their grasp of English is quite rudimentary, sufficient only for processing basic information such as the weather, sports, and erotic visuals. Only a fraction of the Internet users can, to whatever degree, comprehend and produce written or spoken utterances in English. As a result, they are denied of the tremendous opportunities that the Internet has to offer, namely, in-depth information and more serious discussions with the netizens all around the world. To be an effective Internet user one has to be equipped with both receptive and productive abilities in English. Such an expectation, however, appears to be neither practical nor realistic as far as Korea is concerned.

Bok Koh-Ill, a professional Korean writer, thinks otherwise. He came up with a radical idea to make English as an official language of Korea along with Korean. Touching on the sensitive theme of Korean nationalism, Bok argued that "putting 'emotional' nationalism under control is not enough." "As the world is rapidly moving toward a Terrestrial Empire," he further asserted, "not only the political supranational organizations such as the WTO and IMF but also the communication tools like the Internet are increasingly integrated into one unity." Now comes his argument concerning English, the dominant language on the Internet. Bok observed that "the emergence of the universal language is interlocked with the enlargement of network." He predicted that "the Korean language will soon be little more than a museum piece as more nations will adopt a bilingual system in which their own languages coexist with English." His underlying perspective about language is that "language is a tool, and worshipping it as an idol is irrational."

Sensationalism notwithstanding, Bok's audacious proposal immediately touched off a flurry of verbal battle between the conservatives and liberals in Korea. In a wild exchange of high-pitched diatribes, Bok's opponents argued that English was not
panacea and accepting his idea would wither Korea’s unique traditions and hamstring the very cultural foundation.

Bok’s supporters applauded accuracy of his observation that Koreans possess an excessive national pride, which Bok sees as the obstacle to the advancement of the Korean society. Bok even dared all his critics by asking, “If you had to choose between English and Korean for your child’s native language, which one would you select? If you choose English, your child is sure to have faster access to cutting-edge technology and information. If the Korean language is your choice, your child is doomed to lag behind in competition with others.” Whether he will succeed in his effort is far from certain. Bok says, “In our society, nationalism and national language are far too sensitive issues to be discussed with composure.” Yet his other key point—that Koreans are too sensitive to these kinds of issues—seems not that far from reality. A non-random online survey taken immediately after the verbal battle showed 58 percent of those polled were against Bok. Even more noteworthy is the fact that as high as 42 percent sided with Bok and his idea. The deepening dispute seems here to stay, at least for a while.

Multilingual Translation

For a truly multilingual Internet, the long-promised global village, to come of age, there are a host of difficult hurdles to overcome, including, among the many, technical difficulties of communicating in the majority of the world’s languages and development of hardware and software for machine-aided translations. Some progress has been made, and more in sight, in the development of hardware and software for processing texts, from seven-bit ASCII to ISO-Latin, and more recently, to Unicode (ISO 10646), a coding scheme for characters of most of the world’s scripts. Widely hailed as a significant breakthrough in electronic communication worldwide, Unicode character encoding, however, may not be the best one to use in our environment. For example, the Software Laboratory of Nippon Telegraph and Telephone Corporation (NTT) found that Chinese, Japanese, and Korean (CJK) ideographs share the same code space, when Unicode is used in a global search-engine context. Thus, if a Japanese searcher inputs a string for searching, it can equally match against Chinese and Korean counterparts.

Another major problem with Unicode for CJK users is that it doesn’t contain enough code space to capture all ideographs. For the development of a perfect multilingual architecture, incorporating languages that build compound characters or right-to-left ordering needs further tests. Transmitting a message in a language of reader’s preference is one thing; one’s ability to comprehend it is quite another.

In order to send and receive a message, most readers will have to rely on human translators/interpreters. The Internet is home to many language translation sites that offer everything from simple online dictionaries to e-mailed translation services. From the desktop, one can request a translation by selecting to pay for human translation. The global nature of the Internet has proved a boon to translation services, such as TAR Communication in New York, which translated Web-based press releases into 28 languages during the Atlanta Olympic Games in 1996. Translation business via the Internet within the next five years is expected to account for 30 percent of their work.
Given the volume and variety of messages on the Internet, however, exclusive reliance on human translation appears to be an unrealistic proposition. It is too slow and costly to make it a sensible choice for maintaining multilingual websites. A viable alternative is machine-aided translation, which has been vigorously pursued mostly in Japan, Canada and Europe with somewhat mixed results. Systran, available on the French Minitel network since 1983 and Canada’s Meteo system, which translates meteorological bulletins between French and English, are considered to be success stories.

In Korea there are several translation softwares capable of machine-translating foreign language texts into Korean. In an attempt to assess the quality of these softwares, the author had a wide variety of texts in Japanese and English translated into Korean by means of King Sejong, one of the most popular translation softwares in the market. To begin with a conclusion, the Japanese-Korean translation effort proved to be a resounding success, whereas the translated version of the English texts was dismally unsatisfactory.

In terms of the syntax and grammar both Korean and Japanese share a lot of things in common, whereas there is a mile of difference between Korean and English. It seemed as if the intercultural difference involved in the English-Korean translation were an insurmountable wall indeed. The English texts included Abraham Lincoln’s Gettysburg address, a USA Today article headlined “White House Loses Round in Lindsey Case,” a passage in Jean Baudrillard’s La Pensee Radicale, excerpts from the novel Little Prince, and a paragraph in Nico Randeraad’s article “Authority in Search of Liberty.” This mini exercise yielded disconcertingly unsatisfactory results. The translated versions were totally unintelligible, and not good enough to be revisable even with the intervention of human editors. It appeared that the texts replete with sophisticated literary expressions—the Lincoln’s address and Baudrillard’s work in particular—simply defied machine-aided translation.

Of the five different kinds of the English texts, only the USA Today article—concerning a U. S. court’s refusal of the White House appeal to block prosecutors from questioning a presidential aide about Monica Lewinsky—showed some sign of hope. Compared with the other texts, the news article revealed a relatively high level of fidelity in translation. Why did the newspaper article alone fare reasonably well, while other texts simply failed to be translated into Korean? The translator’s familiarity with the subject matter, i.e. President Clinton’s sex scandal, could be an important factor. A more plausible explanation lies in the journalistic style in which the article is written. Journalists are required to adhere to the journalistic principle of brevity and precision when they write about an issue or event.

All the important elements are to be clearly spelled without excessive literary ornamentation. Such a finding could have a practical implication for the information producers on the Internet. For the benefit of a large number of population worldwide they are strongly encouraged to provide a brief and concise abstract of the full text in a machine-friendly manner, so that the abstract can be translated into many languages without losing fidelity. Also useful in multilingual translation will be a two-step process now being implemented by several groups including the United Nations University in
Tokyo. In the two-step process, a text is first thoroughly analyzed into component parts (title, paragraph, sentence), clarified when necessary and possible by a dialogue with the author, then translated into an intermediate, abstract representation—which is used to generate translations in different languages. In this way, readers who have no receptive ability in a foreign language could at least get the gist of the material in hand with the help of translation softwares.

**Human-assisted Machine Translation**

Given the low quality of machine translation and the expensive nature of human translation, the only logical option available to Korea—and for that matter to many countries in the world—appears to be human-assisted/validated machine translation. Experts at numerous regional and international conferences have already addressed to the crucial importance of human-assisted translation. To briefly summarize their recommendations and suggestions:

1) it is vital that a human validator is used to correct automatic translation;

2) it is important to take cultural-specific notions into account when undertaking translations between culturally different linguistic areas;

3) human translations can benefit from processing of files against terminology databases to ensure that technical phrases are translated in a domain-specific way;

4) as keyword matching does not work across domains or across languages, searchable concept-based terminology resources and thesauri are needed;

5) it is important to integrate machine translation and domain-specific terminology sets with authoring tools to speed up translation services.

Technological solution to machine translation is one thing, financial difficulty associated with making machine translation services available through the Internet is another. Because these services take up too many CPU cycles, Information Service Providers would rather offer these services via specialist servers, not as a part of their mainstream operations. The question then arises as to who pays for such servers, and how. Also, the preponderance of English on the Web, with some estimates ranging as high as 95%, appears at odds with the huge investments in translation to be made by the software industry. Investments are likely to be made in the cost-effective European languages, but not in minor languages. Preparing for the multilingual Internet calls for a concerted effort by both public institutions and industry players that produce and utilize language services, tools, and systems.

Regional cooperation among nations that share cultural and linguistic similarities, such as Korea, Japan, and China, must be strongly encouraged not only by the governments concerned but also by regional and international organizations such as UNESCO. Through collaborative R&D arrangements, they can jointly develop multilingual translation tools and services in a much more cost-effective way. As personal computers become less expensive and user-friendlier, people outside North America and European countries are becoming increasingly linked.
In this atmosphere diversity of languages will further enrich the on-line environment by making it possible for people of different cultures and languages to engage themselves in more serious in-depth discussions with other people. There are many advantages of a multilingual Internet: it would allow much more effective and wide diffusion of information and knowledge than would otherwise be possible; common mistakes and misunderstandings resulting from language barrier will be curtailed; and it could be a crack in Americana hegemony over Internet culture.