Journal of Universal Language 3
March 2002, 77-95

The New Babel: Language Barriers on the World Wide Web

Andrew Large
McGill University

Abstract

The relative role of various languages as carriers of web-based information is assessed and the trends examined. Although English is the dominant language, its relative position vis-a-vis other languages is declining as the Web attracts more information from other language communities. The proportion of web users from the English-speaking world is also higher than from any other language group, but their relative strength isdeclining and already they account for less than half of all web users. In this situation of growing language diversity, the language barrierisbecoming stronger. Varioussolutionstothisbarrier are reviewed: multilingual web sites, multilingual approaches by web portals, machine translation and cross language information retrieval systems. Although all theseapproaches can offerpartial remedies, none currently provide a completeanswer. Finally, the rolethat constructed languages can play is assessed.

1. Language Distribution on the Web

Technological developments in computer and communication technologies have by now truly established a virtual global village. As long as the financial resources are available, it is technically feasible to link people by email regardless of where in the world they are located, or to enable someone in one place to create an electronic document that can be read by interested persons anywhere else. The most complete representation of this global village is the World Wide Web that allows the easy exchange of multimedia information in text, images and sound for searching, browsing, viewing and downloading. Organizations and individuals can provide information of all kinds for local, national or international consumption, and in turn can seek such information via the Web. Physical distances have become irrelevant in such electronic communication. To all intents, the office or home on the other side of the world has become as close as the next door office or neighbor.

Although the Web is a product of scientific activity concentrated in one part of the northern hemisphere, it has been adopted with alacrity by the remainder of the world. Indeed, it offers for those regions and countries with relatively poor conventional communication infrastructures and poor collections of traditional printed documents a means to leapfrog into the digital era. Even though access may be more restricted in some parts of the world by the relative expense of computer hardware, the dearth of telephone lines and intermittent electricity supply, few countries now cannot boast of Internet cafes at least in the cities, and a growing number of public and private organizations with their own web sites.

Yet, in this global network of networks, one major barrier continues to impede information communication. The Web might share a common protocol for data transfer, but it does not provide a common natural language for text, sound or image captions. The world might

have become one electronic village, but it is a village inwhich we have no common language of discourse. It is, in effect, the new Babel.

1.1. Web Content

If the earliest days of the Web, at the onset of the 1990s, were dominated by English-language content as well as users, this situation is rapidly changing. In the mercurial world of the Web, where sites appear and disappear overnight and where not even the major web portals individually or collectively can provide access to all the available pages, it is impossible to be precise as to how many individual languages are represented currently on the Web. Nevertheless, it seems probable that information in at least 200 languages is available (Funredes 2001).

Information on the Web is no more distributed equally by language than is information in print, and some languages quantitatively are more strongly represented than others. Undoubtedly, English occupies first place, but its relative dominance appears to be diminishing. In June 1997, the Babel Team, a joint initiative from Alis Technologies (a Montreal-based private company that develops and markets a range of language software tools for the Internet) and the Internet Society, investigated language distribution on the Web (Alis 1997). The Team selected a sample of 3239 home pages drawn from the Web (to be eligible for selection the page had to contain more than 500 characters). The language used on the home page was identified using language analysis software, though it must be emphasised that this software could only identify 17 languages. The Team claim these are the major languages in terms of Web presence. This is seen in Table 1. The Team does point out several potential flaws in the methodology, although it believes that they do not greatly change the picture. One problem is that many bilingual or multilingual sites present the home page in English with hypertext links to other language versions; the software will identify such home pages as being in English and will

ignore the presence of the other hidden language versions. To take account of such flaws, the Team manually verified a small sub-sample of 200 home pages, and extrapolated the resulting analysis corrections to the original sample; these corrections are shown in the final column of Table 1 as "Corrected %". The Babel Team results revealed a dominant position for English; nevertheless, close to two home pages out of ten were in some other language. When the enormous number of pages available on the Web is taken into account--another figure about which no precise estimate can be offered, but certainly well in excess of one billion (for a discussion of the problems of estimating the Web's size, see Dahn 2000)--this nevertheless represents a vast amount of information unavailable in English. German, Japanese, French, and Spanish all accounted for more than one percent.

(1) Table 1: Language Distribution on the Web (June 1997)

Ranking	Language	%	Corrected %
1	English	84.0	82.3
2	German	4.5	4.0
3	Japanese	3.1	1.6
4	French	1.8	1.5
5	Spanish	1.2	1.1
6	Swedish	1.1	0.6
7	Italian	1.0	0.8
8	Portuguese	0.7	0.7
9	Dutch	0.6	0.4
10	Norwegian	0.6	0.3
11	Finnish	0.4	0.3
12	Czech	0.3	0.3
13	Danish	0.3	0.3
14	Russian	0.3	0.1
15	Malay	0.1	0.1
	None/Unknown		5.6
Total			100

Large and Moukdad (2000) analysed the pages indexed by one web portal,AltaVista, in June 1999.AsshowninTable2,theirresultswere comparable with those from the Babel Team; the positions of the top nine languages are identical, and only Chinese, Polish, and Korean from their top 15 are absent from the Babel list (and the automatic language analyser used by Babel was unable to recognize and therefore count pages in Polish or Korean). Although English is used the most, nine other languages are carriers of more than 1.5 million pages each.

(2) Table 2: Language Distribution on the Web (June 1999)

Ranking	Language	Indexed pages
1	English	198,623,158
2	German	20,101,601
3	Japanese	5,265,839
4	French	4,889,844
5	Spanish	4,083,809
6	Swedish	2,539,036
7	Italian	2,523,000
8	Portuguese	2,405,744
9	Dutch	2,346,680
10	Chinese	1,731,619
11	Danish	856,268
12	Czech	852,778
13	Polish	792,194
14	Russian	582,898
15	Korean	566,451

A later analysis was conducted in August 2000 by FUNREDES (2001): the Networks and Development Foundation. It estimates that English accounts for around 60 percent of web pages, a considerable decline from the earlier Babel study. On this basis, it then estimates the percentages for six other European languages, as in Table 3. In all cases, their presence on the Web is proportionately higher than in the Babel study. All but one account for more than two percent of web pages. Furthermore, more than 19 percent of web pages are in languages other than these top seven.

(3) Table 3: Language Distribution on the Web of 7 European Languages (August 2000)

Position	Language	%
1	English	60
2	German	6.3
3	Spanish	4.85
4	French	4.39
5	Italian	2.77
6	Portuguese	2.14
7	Romanian	0.19
	Other	19.36

The difficulties of analyzing content on the Web makes it dangerous to draw firm conclusions about language distribution shifts over time when using figures from several different studies. Grefenstette and Nioche (2000), however, conducted identical (and therefore comparative) studies on the Web in October 1996, August 1999, and February 2000. In these studies, the numbers of words rather than the number of pages have been counted, using the AltaVista search engine. The ratio of seven languages to English is given for each of these three time periods, as illustrated in Table 4. It should be emphasised that in these three studies Grefenstette andNioche only looked at 32languages all of which used the Latin alphabet. Furthermore, they say that AltaVista only indexes about 16 percent of the Web, and it is impossible to know whether the language distribution across the entire Web is reflective of this one portal's indexing coverage.

(4) Table 4: Changes in Language Distribution of 8 European Languages (October 1996-February 2000)

Language	Words Oct 1996	Ratio to English	Words Aug 1999	Ratio to English	Words Feb 2000	Ratio to English
English	6,082,090,000	1.000	28,222,100,000	1.000	48,064,100,000	1.000
German	228,938,428	0.038	1,994,229,409	0.071	3,333,127,671	0.069
French	223,316,023	0.037	1,529,795,169	0.054	2,732,221,327	0.057
Spanish	104,319,158	0.017	1,125,646,460	0.040	1,894,966,981	0.039
Italian	123,555,682	0.020	817,270,444	0.029	1,338,351,674	0.028
Portuguese	106,167,245	0.017	589,391,943	0.021	1,161,898,076	0.024
Norwegian	106,497,066	0.017	669,331,120	0.024	947,486,593	0.020
Finnish	20,647,404	0.003	107,260,274	0.004	166,599,467	0.003

As Table 4 indicates, although English has experienced a growth over the 40 months between the first and the third sampling of 800 percent, itself a dramatic increase, German has grown over the same period by 1500 percent and Spanish by 1800 percent. Only one language, Finnish, failed to register a relative growth compared with English in this period, although even it experienced an enormous absolute growth.

Despite the difficulties of measuring language distribution on the Web, and taking account of the caution required when interpreting the figures presented above, several broad conclusions can be drawn.First, undoubtedly English is the most commonly encountered language on the Web. Second, many other languages, nevertheless, have a presence on the Web. And third, the proportion of English pages is declining compared with those in other languages.

1.2. Web Users

A similar change in language distribution patterns can be identified for web users as for web content. Initially, users were drawn mainly from English-speaking countries: primarily the United States and United Kingdom. More recently, however, language distribution has changed and the proportion of users from other language communities has increased.According to figures from Global Reach (2001), by June 2001, web users from the English-speaking world accounted for 45 percent of the total, while other language groups in total accounted for 55 percent. Japanese speakers were the next most common user population, followed by Chinese and Germans (Table 5). As Web penetration currently is still much lower in the non-English-speaking world than in the English-speaking world, the proportion of web users from the former is likely to rise much higher in the next few years.

(5) Table 5: Web Users by Language (June 2001)

Ranking	Language	% of Web users
1	English	45.0
2	Japanese	9.8
3	Chinese	8.4
4	German	6.2
5	Spanish	5.4
6	Korean	4.7
7	Italian	3.6
8	French	3.4
9	Portuguese	2.5
10	Russian	1.9

The FUNREDES' study (2001) has drawn the interesting conclusion from web content and user statistics that the quantity of web pages produced in a particular language is directly proportional to the number of web users who use that language. This is somewhat surprising as a certain proportion of web pages in English certainly will have been created by non-English speakers; one possible conclusion is that non-English web users are more active in creating web pages than are English speakers.

2. The Language Barrier

The variety of language content and users on the Web is to be welcomed. In particular, it is gratifying to see that content is being made available in local languages for local use; the Web has acquired local as well as national and international significance. Yet, this very language diversity has created a barrier to information retrieval and use.

Information is foundontheWebusingfourtechniques, all of which require linguistic skills on the part of users. First, information can be identified by entering a query to a search engine that will then try to match the query against its indexes of web content. Typically, the query is entered as one or more keywords (normally nouns), linked implicitly (by the search engine), or explicitly (by the user) through Boolean operators. When choosing keywords, it is important to take account of synonyms--failure to do so may result in missed information--and homonyms--failure to do so may result in irrelevant information. The selection of suitable keywords for a search requires a relatively sophisticated command of the language being used, that may prove difficult for someone with partial knowledge of a language, and impossible for someone with little or no knowledge of the language. In a few cases, the search engine is designed to accept a complete sentence rather than keyword queries (a good example is the Ask Jeeves search engine), but even here language proficiency is

required to formulate the most suitable sentence in order to express the information need.

A second way to find information on the Web is to select from hierarchically organized directories (or menus), pioneered by Yahoo! but now offered as an alternative strategy to queries by many search engines. Here, the user is only required to recognize the most suitable directory entry rather than to recall a query. Nevertheless, a level of linguistic competency in the language of the directory entries is required.

Thirdly, the user may browse web content following the hyperlinks liberally added by web page constructors to their pages. Here again, the user must be able to scan the page quickly and select suitable links to follow based on an understanding of the page's content if browsing is to be efficient and successful.

Only the fourth information-seeking strategy is free from linguistic demands on the part of the user; here, the user enters the Uniform Resource Locator (URL) of a required page to go straight to it. Such a strategy can only be used, however, if the URL earlier has been discovered in some way, and relying upon known URLs to find information on the Web is not an effective way to exploit fully its potential.

Successful information retrieval poses challenges to many users even when they are working exclusively in their own language. The size and subject diversity of the Web has produced an environment that is especially challenging. These challenges are, then, further magnified when seeking information in a foreign language.

Once pages have been found, they must be assessed for relevancy--this involves scanning their content--and ultimately they must be read. Here, the barriers to access created by language diversity are similar to those found in more traditional printed information sources (Large 1983).

Although this article will not explore the issue further, digital information raises another set of problems relating to script multiplicity.

Inputting, searching, displaying and printing problems can all occur when switching from one script to another, although various software solutions can eliminate most problems (Large & Moukdad 2000).

3. Solutions

3.1. Web-based solutions

Some relief from the language barrier is offered by content providers. In many cases, web pages are provided in two or more languages, with language selection being made from the home page of the site. Such a policy obviously increases the possibility both of finding and reading the information, as long as users can cope with at least one of the languages employed by the site. It should be noted, however, that content is not always identical in all the language versions.

Web portals, search engines and/or directories, may also offer a measure of linguistic flexibility. AltaVista has long set an example of providing multilingual options. Ithasestablished national versions in 22 countries located in Asia, Australasia, Europe, North and South America. When one is chosen the AltaVista interface changes to the language primarily used in that country. AltaVista also offers the opportunity to search in 25 different languages as well as in all languages together. To take another example, Yahoo! also has individual access sites in 22 countries, and in some cases offers search options in the country's own language (Spanish in the case of Mexico, for example) or languages (English or French in the case of Canada).

A growing number of specialised web portals have been developed to serve individual countries, such as EgyptSearch.Com (www.egyptserach.com) for Egypt, Maple Square (www.maplesquare.ca) for Canada, and Max (www.max.co.za) for South Africa. Other portals offer interfaces and searching in a specific language or languages;

examples are Phantis (www.phantis.com) in Greek or English, Rambler (www.rambler.ru) in Russian, Fireball (www.fireball.de) in German, and Surfboard (surfboard.ixquick.com) in Dutch.

Such approaches are to be welcomed, and clearly lower language barriers for some language communities. Nevertheless, they only offer a very partial solution to universal web access.

3.2. Technological Solutions

Technological developments offer two related palliatives to the access problems posed by linguistic diversity on the Web: machine translation (henceforth MT) and cross language information retrieval (henceforth CLIR).

MT has a long and tempestuous history dating back to the 1950s. The task of programming a computer to translate from one language into another language has proved much more difficult than the early pioneers anticipated. Nevertheless, progress has been accomplished and many commercial MT systems are now operating across a growing number of languages (see Hutchins & Somers 1992 for an overview of MT and Maegaard 1999 for a brief review of the current situation).

A number of MT systems are available on the Web and several are free of charge. The best known is Babel Fish, available from Alta-Vista's home page. It can be used either to translate short text extracts by entering them in the translation box provided, or to translate web pages by entering the page's URL. Currently, it is available from English into eight languages (Chinese, French, German, Italian, Japanese, Korean, Portuguese and Spanish); from German into English or French; from French into English or German; and from Chinese, Italian, Japanese, Korean, Portuguese, Russian and Spanish into English.

Unfortunately, the diversity of content on the Web poses special challenges to MT systems. The necessity for enormous dictionaries and the likelihood of problems arising from attempting to translate a

synonym in one language into its correct meaning in the second language (a major and long-standing complication for MT) often produces very low quality translations, in some cases largely unintelligible. An additional drawback is that the number of language pairs available, though expanding, is still very limited. Nevertheless, such MT systems can provide free and fast rough translations for some web users.

CLIR attempts to solve the problem of information seeking across language boundaries (Oard 1998). Using CLIR, information seekers enter a query to a search engine in one language (the source language or SL) in order to search for information in a second language (the target language or TL). The CLIR system translates the SL query into the TL. The retrieved documents are then displayed to the user. In some cases they will first be translated from the TL into the SL using an MT system, but in other cases they are displayed in the TL. The assumption here is that the user has sufficient familiarity with the TL to make decisions about relevance or even to read and understand the information, but insufficient familiarity to formulate the initial queries in it. A few CLIR systems have also experimented with various display options to enable users to make decisions about the relevance of retrieved pages in the TL even though they cannot understand it. For example, thumbnail views of the retrieved page can be displayed showing the location in that page of the query words (in the TL); if the query words occur in the title or opening paragraph, or occur in close proximity to each other, for instance, the page is likely to prove more relevant (when translated) than if the query words are scattered through the text. CLIR, then, uses MT techniques, but in a special context.

A number of working CLIR systems are available off the Web, but they mainly operate with controlled language rather than natural language retrieval systems. In retrieval systems using a controlled language, the query terms are chosen from a controlled list of terms where one word is used to represent one concept, and where one

concept is always expressed by one word only. In other words, the problems of synonyms and homonyms that plague natural language retrieval are circumvented by using an artificial indexing language. The documents to be retrieved are then indexed using terms selected from this same controlled list. In order to provide CLIR, it is, then, necessary to translate the controlled term list from the SL into the TL. If this is done well, there is no reason why the CLIR system should operate any less effectively than its monolingual equivalent. Unfortunately, in the anarchic world of the Web the use by all information providers of a controlled vocabulary to describe content and the use byallinformation seekers of the same vocabulary is impossible. On the Web, most searching must take place using natural and not controlled language.

Several experimental CLIR systems are available on the Web. ARCTOS (messene.nmsu.edu/ursa/arctos/), from New Mexico State University, uses a combination of automatic and user-assisted methods to build and improve cross-language queries.MUNDIAL (crl.nmsu.edu/Research/Projects/tipster/ursa/Mundial/mundial.html), also from New Mexico State University,searches for web pages in multiple languages given an initial query in English. Currently, CLIR systems only function between a few language pairs, the translations are not always successful, and they cannot assist with hypertext browsing. Research is active in this field, however, and considerable progress is likely to be made before too long.

4. Constructed Languages and the Web

Constructed languages long have offered a solution to the problems of language multiplicity for international communication (Large 1985). In the multilingual environment of the Web, do they offer an answer to information seeking across language boundaries?

The Web certainly has become home to many pages written in and about the various constructed languages. A search in early July 2001

on AltaVista using the query keyword *Esperanto*, for example, retrieved 542,607 pages! The Web has become yet another means for supporters of constructed languages to organize themselves, communicate between themselves and inform the world about themselves. It is an effective way to find about these languages, and a substantial amount of reference material relating to the languages can be found on the Web. Finding such information is helped by electronic bibliographical tools such as the Virtuala Esperanto: Biblioteko (www.esperanto.net/veb/).

Constructed languages are also considered sufficiently important for a web portal such as Yahoo! to offer a route to relevant material via its hierarchically arranged directories. Following the directory, path from the main heading Social Science to Languages, and then Constructed Languages will lead to Fictional Languages and IALs. Under this, last heading are separate entries for Esperanto, Eurolang, Ido, and Interlingua, each leading to many pages on these languages. At least one multi-language web portal, Euroseek.com (www.euroseek.com), offers Esperanto as one of its 29 languages to which searches can be confined. It also allows users to select Esperanto as the language of its home page interface.

Might a constructed language become the digital IAL forweb-based communication? Grefenstette and Nioche (2000) list Esperanto among the languages on the Web that employ the Latin alphabet for which they estimated the number of words. But inFebruary 2000, it occupied only 27th position out of 32 languages, with 26,795,000 individual words (not pages), above Latvian, Lithuanian, Breton, Albanian, and Welsh, but below such languages as Latin (in 25th place), Basque, Irish, and Estonian. Its relative impact is, therefore, small (though greater than any other constructed language).

5. The Web of the Future

Assuming that the Web continues to thrive as an international medium for information exchange (by no means certain in the rapidly evolving digital landscape), is language likely to prove a continuing barrier? One scenario would be for English to become de facto the language of the Web. It is already by far the most commonly used language, accounting for more pages than any other. The quantitative data already presented above, however, suggests that the proportion of web pages in English is declining; although the total number of pages in English continues to increase, the number of pages in other languages is growing even more quickly. The same can be said of web users; numbers from the English-speaking countries are growing more slowly than those from elsewhere, and the former are already in a minority. Furthermore, the vast size of the web means that even small percentages in any one language translate into very large numbers of pages. Unless the trend against English is reversed, then, it is not going to provide the solution to the language barrier.

Growing numbers of web sites offering bilingual or multilingual access greatly help their users. However, such an approach can never offer more than a partial solution to the language barrier. Such sites are always likely to remain a tiny minority (although they may be among the more heavily visited) because they require time effort and linguistic skills to produce. In any case, no site can offer access via all the world's languages, and in practice most sites adopting this practice are in the local language plus English and perhaps one or two other major languages. The availability of web portals in various languages as well as specialist web portals for individual language communities helps to make information in these languages more available to those familiar with them, but does little to help information seekers from other communities gain similar access.

The Web is not only a source of textual information, but also of

still images, animation sequences, video clips, and sound. Visual information and sound effects do bypass the problems of language (though not necessarily of variations in interpretation by different cultural groups). But such non-textual sources of information remain secondary to text on the Web, and in any case they must often be retrieved by searching on text in captions and the like. Voice recognition systems for limited applications are now available, and research is progressing on systems that might be adopted in broader environments such as the Web (Haynes 1998). The replacement of textual by aural information, however, would not eliminate language problems but merely change their nature (Mariani 1999).

The technological solutions offered by machine translation and cross language information retrieval are likely to become more important in future. Yet, the challenges are considerable and the numbers of languages currently involved rather small. For the immediate future their impact is likely to remain limited.

The digitization of information and its availability via high speed networks has not in any way reduced the potential contribution of constructed languages to effective, equal and unbiased international communication. In this domain as in the earlier print-based culture, however, their many virtues notwithstanding, constructed languages do not seem poised to offer a real solution. The supporters of various constructed languages seem to be using the Web effectively for their own purposes, but as far as any language emerging as the world's international digital language, there is little cause for optimism.

The Web is likely to continue as an exciting place for international communication, with an increasingly rich and varied collection of information from all parts of the world. But short of users learning more languages the barrier to full access created by language diversity will persist.

References

Alis Technologies. 1997. Web Languages Hit Parade. Available at URL <http://alis.isoc.org/palmares.en.html>.

Dahn, M. 2000. Counting Angels on a Pinhead: Critically Interpreting Web Size Estimates. Online 24 (1).Available at URL <http://www.onlineinc.com/onlinemag/OL2000/dahn1.html>.

Funredes. 2001. Languages and Cultures. Available at URL <http://funredes.org/LC/english>.

Global Reach. 2001.Global Internet Statistics. Available at URL <http://www.glreach/com/eng/index.php3>.

Grefenstette, G. & J. Nioche. 2000. Estimation of English and Non-English Language Use on the WWW. Available at URL <http://citeseer.nj.nec.com/410377.html>.

Haynes, C. 1998. Translating and Interpreting by Computer. *Managing Information* 5.1, 38-39.

Hutchins, W. J. & H. L. Somers. 1992. *An Introduction to Machine Translation*. London: Academic Press.

Large, A. 1983. *The Foreign-language Barrier: Problems in Scientific Communication.* London: Deutsch.

Large, A. 1985. *The Artificial Language Movement*. Oxford: Blackwell.

Large, A. & H. Moukdad. 2000. Multilingual Access to Web Resources: An Overview. *Program* 34.1, 43-58.

Maegaard, B. 1999. Machine Translation. Available at URL <http://www.cs.cmu.edu/~ref/mlim/chapter4.html>.

Mariani, J. 1999. Multilingual Speech Processing (Recognition and Synthesis). Available at URL <http://www.cs.cmu.edu/~ref/mlim/chapter5.html>.

Oard, D. 1998. Cross-language Information Retrieval. In M.E.Williams(ed.), *Annual Review of Information Science and Technology* 33, 223-256. Medford: Information Today.