THE INTERNET AND LANGUAGES

[around the year 2000]

MARIE LEBERT

NEF, University of Toronto, 2009

TABLE

INTRODUCTION

It is true that the internet transcends the limitations of time, distances and borders, but what about languages? Non-English-speaking internet users reached 50% in July 2000.

"Language Nations"

"Because the internet has no national boundaries, the organization of users is bounded by other criteria driven by the medium itself. In terms of multilingualism, you have virtual communities, for example, of what I call 'Language Nations'… all those people on the internet wherever they may be, for whom a given language is their native language. Thus, the Spanish Language nation includes not only Spanish and Latin American users, but millions of Hispanic users in the U.S., as well as odd places like Spanish-speaking Morocco." (Randy Hobler, consultant in internet marketing for translation products and services, September 1998)

"Linguistic Democracy"

"Whereas 'mother-tongue education' was deemed a human right for every child in the world by a UNESCO report in the early 1950s, 'mother- tongue surfing' may very well be the Information Age equivalent. If the internet is to truly become the Global Network that it is promoted as being, then all users, regardless of language background, should have access to it. To keep the internet as the preserve of those who, by historical accident, practical necessity, or political privilege, happen to know English, is unfair to those who don't." (Brian King, director of the WorldWide Language Institute, September 1998)

A medium for the world

"It is very important to be able to communicate in various languages. I would even say this is mandatory, because the information given on the internet is meant for the whole world, so why wouldn't we get this information in our language or in the language we wish? Worldwide information, but no broad choice for languages, this would be quite a contradiction, wouldn't it?" (Maria Victoria Marinetti, teacher in Spanish and translator, August 1999)

Good software

"When software gets good enough for people to chat or talk on the web in real time in different languages, then we will see a whole new world appear before us. Scientists, political activists, businesses and many more groups will be able to communicate immediately without having to go through mediators or translators." (Tim McKenna, writer and philosopher, October 2000)

***

Unless specified otherwise, quotations are excerpts from NEF interviews. Many thanks to all those who are quoted in this book, and who kindly answered questions about multilingualism over the years. Most interviews are available online <http://www.etudes- francaises.net/entretiens/>. This book is also available in French, with a different text. Both versions are available online <http://www.etudes-francaises.net/entretiens/multi.htm>. The author, whose mother tongue is French, is responsible for any remaining mistakes in English.

Marie Lebert is a researcher and editor specializing in technology for books, other media, and languages. Her books are published by NEF (Net des études françaises / Net of French Studies), University of Toronto, Canada, and are freely available online <http://www.etudes- francaises.net>.

"LANGUAGE NATIONS" ONLINE

= [Quote]

Randy Hobler, a consultant in internet marketing for Globalink, a company specializing in language translation software and services, wrote in September 1998: "Because the internet has no national boundaries, the organization of users is bounded by other criteria driven by the medium itself. In terms of multilingualism, you have virtual communities, for example, of what I call 'Language Nations'… all those people on the internet wherever they may be, for whom a given language is their native language. Thus, the Spanish Language nation includes not only Spanish and Latin American users, but millions of Hispanic users in the U.S., as well as odd places like Spanish-speaking Morocco."

= [Text]

At first, the internet was nearly 100% English. A network was set up by the Pentagon in 1969, before spreading to U.S. governmental agencies and universities from 1974 onwards, after Vinton Cerf and Bob Kahn invented TCP/IP (transmission control protocol / internet protocol). After the creation of the World Wide Web in 1989-90 by Tim Berners-Lee at the European Laboratory for Particle Physics (CERN) in Geneva, Switzerland, and the distribution of the first browser Mosaic, the ancestor of Netscape, from November 1993 onwards, the internet really took off, first in the U.S. and Canada, then worldwide.

Why did the internet spread in North America first? The U.S. and Canada were leading the way in computer science and communication technology, and a connection to the internet, mainly through a phone line at the time, was much cheaper than in most countries. In Europe, avid internet users needed to navigate the web at night, when phone rates by the minute were cheaper, to cut their expenses. In 1998, some French, Italian and German users were so fed up with the high rates that they launched a movement to boycott the internet one day per week, for internet providers and phone companies to set up a special monthly rate for them. This paid off, and providers began to offer monthly "internet rates".

In the 1990s, the percentage of English decreased from nearly 100% to 80%. People from all over the world began to have access to the internet, and to post more and more webpages in their own languages.

The first major study about language distribution on the web was run by Babel, a joint initiative from Alis Technologies, a company specializing in language translation services, and the Internet Society. The results were published in June 1997 on a webpage named "Web Languages Hit Parade". The main languages were English with 82.3%, German with 4.0%, Japanese with 1.6%, French with 1.5%, Spanish with 1.1%, Swedish with 1.1%, and Italian with 1.0%.

In "Web Embraces Language Translation", an article published in ZDNN (ZDNetwork News) on 21 July 1998, Martha L. Stone explained: "This year, the number of new non-English websites is expected to outpace the growth of new sites in English, as the cyber world truly becomes a 'World Wide Web'."

According to Global Reach, a branch of Euro-Marketing Associates, an international marketing consultancy, there were 56 million non-English- speaking users in July 1998, with 22.4% Spanish-speaking users, 12.3% Japanese-speaking users, 14% German-speaking users, and 10% French-speaking users. But 80% of all webpages were still in English, whereas only 6% of the world population was speaking English as a native language, while 16% was speaking Spanish as a native language. 15% of Europe's half a billion population spoke English as a first language, 28% didn't speak English at all, and 32% were using the web in English.

Jean-Pierre Cloutier was the editor of "Chroniques de Cybérie", a weekly French-language online report of internet news. He wrote in August 1999: "We passed a milestone this summer. Now more than half the users of the internet live outside the United States. Next year, more than half of all users will be non English-speaking, compared with only 5% five years ago. Isn't that great? (...) The web is going to grow in non-English-speaking regions. So we have to take into account the technical aspects of the medium if we want to reach these 'new' users. I think it is a pity there are so few translations of important documents and essays published on the web - from English into other languages and vice versa. (...) In the same way, the recent spreading of the internet in new regions raises questions which would be good to read about. When will Spanish-speaking communication theorists and those speaking other languages be translated?"

Will the web hold as many languages as the ones spoken on our planet? This will be quite a challenge, with the 6,700 languages listed in "The Ethnologue: Languages of the World", an authoritative catalog published by SIL International (SIL: Summer Institute of Linguistics) and freely available on the web since the mid-1990s.

The year 2000 was a turning point for a multilingual internet, regarding its users. Non English-speaking users reached 50% in summer 2000. According to Global Reach, they were 52.5% in summer 2001, 57% in December 2001, 59.8% in April 2002, 64.4% in September 2003 (including 34.9% non-English-speaking Europeans and 29.4% Asians), and 64.2% in March 2004 (including 37.9% non-English-speaking Europeans and 33% Asians).

Despite the so-called English-language hegemony some non-English- speaking intellectuals were complaining about, without doing much to promote their own language, the internet was also a good medium for minority languages, as stated by Caoimhín Ó Donnaíle. Caoimhín has taught computing at the Institute Sabhal Mór Ostaig, on the Island of Skye (Scotland). He has also created and maintained the college website, as the main site worldwide with information on Scottish Gaelic, with a bilingual (English, Gaelic) list of European minority languages. He wrote in May 2001: "Students do everything by computer, use Gaelic spell-checking, a Gaelic online terminology database. There are more hits on

our website. There is more use of sound. Gaelic radio (both Scottish and Irish) is now available continuously worldwide via the internet. A major project has been the translation of the Opera web-browser into Gaelic - the first software of this size available in Gaelic."

TOWARDS A "LINGUISTIC DEMOCRACY"

= [Quote]

Brian King, director of the WorldWide Language Institute (WWLI), brought up the concept of "linguistic democracy" in September 1998: "Whereas 'mother-tongue education' was deemed a human right for every child in the world by a UNESCO report in the early 1950s, 'mother- tongue surfing' may very well be the Information Age equivalent. If the internet is to truly become the Global Network that it is promoted as being, then all users, regardless of language background, should have access to it. To keep the internet as the preserve of those who, by historical accident, practical necessity, or political privilege, happen to know English, is unfair to those who don't."

= [Text]

Yoshi Mikami, a computer scientist at Asia Info Network in Fujisawa (Japan), launched in December 1995 the website "The Languages of the World by Computers and the Internet", also known as the Logos Home Page or Kotoba Home Page. (The website was updated until September 2001.) Yoshi was also the co-author (with Kenji Sekine and Nobutoshi Kohara) of "The Multilingual Web Guide" (Japanese edition), a print book published by O'Reilly Japan in August 1997, and translated in 1998 into English, French and German.

Yoshi Mikami explained in December 1998: "My native tongue is Japanese. Because I had my graduate education in the U.S. and worked in the computer business, I became bilingual in Japanese and American English. I was always interested in languages and different cultures, so I learned some Russian, French and Chinese along the way. In late 1995, I created on the web 'The Languages of the World by Computers and the Internet' and tried to summarize there the brief history, linguistic and phonetic features, writing system and computer processing aspects for each of the six major languages of the world, in English and Japanese. As I gained more experience, I invited my two associates to help me write a book on viewing, understanding and creating multilingual webpages, which was published in August 1997 as 'The Multilingual Web Guide', in a Japanese edition, the world's first book on such a subject."

Yoshi added in the same email interview: "Thousands of years ago, in Egypt, China and elsewhere, people were more concerned about communicating their laws and thoughts not in just one language, but in several. In our modern world, most nation states have each adopted one language for their own use. I predict greater use of different languages and multilingual pages on the internet, not a simple gravitation to American English, and also more creative use of multilingual computer translation. 99% of the websites created in Japan are written in Japanese."

Robert Ware launched his website OneLook Dictionaries in April 1996 as a "fast finder" in hundreds of online dictionaries. On September 2, 1998, the fast finder could "browse" 2,058,544 words in 425 dictionaries covering various topics: business, computer/internet, medical, miscellaneous, religion, science, sports, technology, general, and slang. OneLook Dictionaries was provided as a free service by the company Study Technologies, in Englewood, Colorado.

Robert Ware explained in September 1998: "On the personal side, I was almost entirely in contact with people who spoke one language and did not have much incentive to expand language abilities. Being in contact with the entire world has a way of changing that. And changing it for the better! (...) I have been slow to start including non-English dictionaries (partly because I am monolingual). But you will now find a few included."

In the same email interview, Robert wrote about a personal experience showing the internet could promote both a common language and multilingualism: "In 1994, I was working for a college and trying to install a software package on a particular type of computer. I located a person who was working on the same problem and we began exchanging email. Suddenly, it hit me... the software was written only 30 miles away but I was getting help from a person half way around the world. Distance and geography no longer mattered! OK, this is great! But what is it leading to? I am only able to communicate in English but, fortunately, the other person could use English as well as German which was his mother tongue. The internet has removed one barrier (distance) but with that comes the barrier of language. It seems that the internet is moving people in two quite different directions at the same time. The internet

(initially based on English) is connecting people all around the world. This is further promoting a common language for people to use for communication. But it is also creating contact between people of different languages and creates a greater interest in multilingualism. A common language is great but in no way replaces this need. So the internet promotes both a common language *and* multilingualism. The good news is that it helps provide solutions. The increased interest and need is creating incentives for people around the world to create improved language courses and other assistance, and the internet is providing fast and inexpensive opportunities to make them available."

The internet could also be a tool to develop a "cultural identity". During the Symposium on Multimedia Convergence organized by the International Labor Office (ILO) in January 1997, Shinji Matsumoto, general secretary of the Musicians' Union of Japan (MUJ), explained: "Japan is quite receptive to foreign culture and foreign technology. (…) Foreign culture is pouring into Japan and, in fact, the domestic market is being dominated by foreign products. Despite this, when it comes to preserving and further developing Japanese culture, there has been insufficient support from the government. (…) With the development of information networks, the earth is getting smaller and it is wonderful to be able to make cultural exchanges across vast distances and to deepen mutual understanding among people. We have to remember to respect national cultures and social systems."

December 1997 was a turning point for a plurilingual web. AltaVista, a leading search engine, was the first website to launch a free translation software called Babel Fish (or AltaVista Translation), which could translate up to three pages from English into French, German, Italian, Portuguese or Spanish, and vice versa. Non-English- speaking users were thrilled. The software was developed by Systran, a pioneer company specializing in machine translation. Later on, other translation software was developed by Alis Technologies, Globalink, Lernout & Hauspie, Softissimo, Wordfast and Trados, with free and/or paid versions available on the web.

Brian King, director of the WorldWide Language Institute (WWLI), brought up the concept of "linguistic democracy" in September 1998: "Whereas 'mother-tongue education' was deemed a human right for every child in the world by a UNESCO report in the early 1950s, 'mother- tongue surfing' may very well be the Information Age equivalent. If the internet is to truly become the Global Network that it is promoted as being, then all users, regardless of language background, should have access to it. To keep the internet as the preserve of those who, by historical accident, practical necessity, or political privilege, happen to know English, is unfair to those who don't."

Geoffrey Kingscott was the managing director of Praetorius, a language consultancy in applied languages. He wrote in September 1998: "Because the salient characteristics of the web are the multiplicity of site generators and the cheapness of message generation, as the web matures it will in fact promote multilingualism. The fact that the web originated in the USA means that it is still predominantly in English but this is only a temporary phenomenon. If I may explain this further, when we relied on the print and audiovisual (film, television, radio, video, cassettes) media, we had to depend on the information or entertainment we wanted to receive being brought to us by agents (publishers, television and radio stations, cassette and video producers) who have to subsist in a commercial world or — as in the case of public service broadcasting — under severe budgetary restraints. That means that the size of the customer-base is all- important, and determines the degree to which languages other than the ubiquitous English can be accommodated. These constraints disappear with the web. To give only a minor example from our own experience, we publish the print version of Language Today [a magazine for linguists, published by Praetorius] only in English, the common denominator of our readers. When we use an article which was originally in a language other than English, or report an interview which was conducted in a language other than English, we translate into English and publish only the English version. This is because the number of pages we can print is constrained, governed by our customer-base (advertisers and subscribers). But for our web edition we also give the original version."

Founder of Euro-Marketing Associates and its virtual branch Global Reach, Bill Dunlap was championing the assets of e-commerce in Europe among his fellow compatriots in the U.S. Bill wrote in December 1998: "There are so few people in the U.S. interested in communicating in many languages — most Americans are still under the delusion that the rest of the world speaks English. However, here in Europe (I'm writing from France), the countries are small enough so that an international perspective has been necessary for centuries."

As the internet quickly spread worldwide, more and more people in the U.S. realized that, although English may stay the main international language for exchanges of all kinds, people did prefer to read information in their own language. To reach as large an audience as possible, companies and organizations needed to offer bilingual, trilingual, even multilingual websites, while adapting their content to a given audience. Thus the need of both localization and internationalization, which became a major trend in the following years, not only in the U.S. but in many countries, with companies setting

up bilingual websites, in their language and in English, to reach a wider audience, and get more clients.

Brian King, director of the WorldWide Language Institute (WWLI), explained in September 1998: "As well as the appropriate technology being available so that the non-English speaker can go, there is the impact of 'electronic commerce' as a major force that may make multilingualism the most natural path for cyberspace. A pull from non- English-speaking computer users and a push from technology companies competing for global markets has made localization a fast growing area in software and hardware development."

In 1998, the European Network in Language and Speech (ELSNET) was a network of more than 100 European academic and industrial institutions. ELSNET members intended to build multilingual speech and natural language systems with coverage of both spoken and written language. Steven Krauwer, coordinator of ELSNET, explained in September 1998: "As a European citizen I think that multilingualism on the web is absolutely essential, as in the long run I don't think that it is a healthy situation when only those who have a reasonable command of English can fully exploit the benefits of the web. As a researcher (specialized in machine translation) I see multilingualism as a major challenge: how can we ensure that all information on the web is accessible to everybody, irrespective of language differences."

Steven added in August 1999: "I've become more and more convinced we should be careful not to address the multilinguality problem in isolation. I've just returned from a wonderful summer vacation in France, and even if my knowledge of French is modest (to put it mildly), it's surprising to see that I still manage to communicate successfully by combining my poor French with gestures, facial expressions, visual clues and diagrams. I think the web (as opposed to old-fashioned text-only email) offers excellent opportunities to exploit the fact that transmission of information via different channels (or modalities) can still work, even if the process is only partially successful for each of the channels in isolation."

What practical solutions would he suggest for a truly multilingual web? "At the author end: better education of web authors to use combinations of modalities to make communication more effective across language barriers (and not just for cosmetic reasons). At the server end: more translation facilities à la AltaVista (quality not impressive, but always better than nothing). At the browser end: more integrated translation facilities (especially for the smaller languages), and more quick integrated dictionary lookup facilities."

Linguistic pluralism and diversity are everybody's business, as explained in a petition launched by the European Committee for the Respect of Cultures and Languages in Europe (ECRCLE) "for a humanist and multilingual Europe, rich of its cultural diversity": "Linguistic pluralism and diversity are not obstacles to the free circulation of men, ideas, goods and services, as would like to suggest some objective allies, consciously or not, of the dominant language and culture. Indeed, standardization and hegemony are the obstacles to the free blossoming of individuals, societies and the information economy, the main source of tomorrow's jobs. On the contrary, the respect for languages is the last hope for Europe to get closer to the citizens, an objective always claimed and almost never put into practice. The Union must therefore give up privileging the language of one group." The full text of the petition was available in the eleven official languages of the European Union. Among other things, the petition asked the revisors of the Treaty of the European Union to include the respect of national cultures and languages in the text of the treaty, and the national governments to "teach the youth at least two, and preferably three foreign European languages; encourage the national audiovisual and musical industries; and favour the diffusion of European works."

Henk Slettenhaar is a professor in communication technology at Webster University in Geneva, Switzerland. Henk is a trilingual European. He is Dutch, he teaches computer science in English, and he is fluent in French as a resident in neighboring France. He has regularly insisted on the need of bilingual websites, in the original language and in English. He wrote in December 1998: "I see multilingualism as a very important issue. Local communities which are on the web should use the local language first and foremost for their information. If they want to be able to present their information to the world community as well, their information should be in English as well. I see a real need for bilingual websites. (...) As far as languages are concerned, I am delighted that there are so many offerings in the original languages now. I much prefer to read the original with difficulty than to get a bad translation."

Henk added in August 1999: "There are two main categories of websites in my opinion. The first one is the global outreach for business and information. Here the language is definitely English first, with local versions where appropriate. The second one is local information of all kinds in the most remote places. If the information is meant for people of an ethnic and/or language group, it should be in that language first, with perhaps a summary in English. We have seen lately how important these local websites are — in Kosovo and Turkey, to mention just the most recent ones. People were able to get

information about their relatives through these sites."

Marcel Grangier was the head of the French Section of the Swiss Federal Government's Central Linguistic Services, which means he was in charge of organizing translations into French for the Swiss government. He wrote in January 1999: "We can see multilingualism on the internet as a happy and irreversible inevitability. So we have to laugh at the doomsayers who only complain about the supremacy of English. Such supremacy is not wrong in itself, because it is mainly based on statistics (more PCs per inhabitant, more people speaking English, etc.). The answer is not to 'fight' English, much less whine about it, but to build more sites in other languages. As a translation service, we also recommend that websites be multilingual. The increasing number of languages on the internet is inevitable and can only boost multicultural exchanges. For this to happen in the best possible circumstances, we still need to develop tools to improve compatibility. Fully coping with accents and other characters is only one example of what can be done."

Alain Bron, a consultant in information systems and a writer, wrote in January 1999: "Different languages will still be used for a long time to come and this is healthy for the right to be different. The risk is of course an invasion of one language to the detriment of others, and with it the risk of cultural standardization. I think online services will gradually emerge to get around this problem. First, translators will be able to translate and comment on texts by request, but mainly sites with a large audience will provide different language versions, just as the audiovisual industry does now."

Guy Antoine, founder of Windows on Haiti, a reference website about Haitian culture, wrote in November 1999: "It is true that for all intents and purposes English will continue to dominate the web. This is not so bad in my view, in spite of regional sentiments to the contrary, because we do need a common language to foster communications between people the world over. That being said, I do not adopt the doomsday view that other languages will just roll over in submission. Quite the contrary. The internet can serve, first of all, as a repository of useful information on minority languages that might otherwise vanish without leaving a trace. Beyond that, I believe that it provides an incentive for people to learn languages associated with the cultures about which they are attempting to gather information. One soon realizes that the language of a people is an essential and inextricable part of its culture. (…)

From this standpoint, I have much less faith in mechanized tools of language translation, which render words and phrases but do a poor job of conveying the soul of a people. Who are the Haitian people, for instance, without "Kreyòl" (Creole for the non-initiated), the language that has evolved and bound various African tribes transplanted in Haiti during the slavery period? It is the most palpable exponent of commonality that defines us as a people. However, it is primarily a spoken language, not a widely written one. I see the web changing this situation more so than any traditional means of language dissemination. In Windows on Haiti, the primary language of the site is English, but one will equally find a center of lively discussion conducted in "Kreyòl". In addition, one will find documents related to Haiti in French, in the old colonial creole, and I am open to publishing others in Spanish and other languages. I do not offer any sort of translation, but multilingualism is alive and well at the site, and I predict that this will increasingly become the norm throughout the web."

ENCODING: FROM ASCII TO UNICODE

= [Quote]

Brian King, director of the WorldWide Language Institute (WWLI), explained in September 1998: "The first step was for ASCII to become Extended ASCII. This meant that computers could begin to start recognizing the accents and symbols used in variants of the English alphabet — mostly used by European languages. But only one language could be displayed on a page at a time. (…) The most recent development is Unicode. Although still evolving and only just being incorporated into the latest software, this new coding system translates each character into 16 bytes. Whereas 8-byte extended ASCII could only handle a maximum of 256 characters, Unicode can handle over 65,000 unique characters and therefore potentially accommodate all of the world's writing systems on the computer. So now the tools are more or less in place. They are still not perfect, but at last we can at least surf the web in Chinese, Japanese, Korean, and numerous other languages that don't use the Western alphabet. As the internet spreads to parts of the world where English is rarely used - such as China, for example, it is natural that Chinese, and not English, will be the preferred choice for interacting with it. For the majority of the users in China, their mother tongue will be the only choice."

= Encoding in Project Gutenberg

Used since the beginning of computing, ASCII (American Standard Code for Information Interchange) is a 7-bit coded character set for information interchange in English. It was published in 1968 by ANSI (American National Standards Institute), with an update in 1977 and 1986. The 7-bit plain ASCII, also called Plain Vanilla ASCII, is a set of 128 characters with 95 printable unaccented characters (A-Z, a-z, numbers, punctuation and basic symbols), i.e. the ones that are available on the English/American keyboard. With the use of other European languages, extensions of ASCII (also called ISO-8859 or ISO-Latin) were created as sets of 256 characters to add accented characters as found in French, Spanish and German, for example ISO 8859-1 (ISO-Latin-1) for French.

Created by Michael Hart in July 1971, Project Gutenberg was the first information provider on the internet. Michael's purpose was to digitize as many literary texts as possible, and to offer them for free in a digital library open to anyone. Michael explained in August 1998: "We consider etext to be a new medium, with no real relationship to paper, other than presenting the same material, but I don't see how paper can possibly compete once people each find their own comfortable way to etexts, especially in schools."

Whether digitized years ago or now, all Project Gutenberg books are created in 7-bit plain ASCII, called Plain Vanilla ASCII. When 8-bit ASCII is used for books with accented characters like French or German, Project Gutenberg also produces a 7-bit ASCII version with the accents stripped. (This doesn't apply for languages that are not "convertible" in ASCII, like Chinese, encoded in Big-5.)

Project Gutenberg sees Plain Vanilla ASCII as the best format by far, and calls it "the lowest common denominator". It can be read, written, copied and printed by any simple text editor or word processor on any electronic device. It is the only format compatible with 99% of hardware and software. It can be used as it is or to create versions in many other formats. It will still be used while other formats will be obsolete, or are already obsolete, like formats of a few short-lived reading devices launched since 1999. It is the assurance collections will never be obsolete, and will survive future technological changes. The goal is to preserve the texts not only over decades but over centuries.

Project Gutenberg also publishes ebooks in well-known formats like HTML, XML or RTF. There are Unicode files too. Any other format provided by volunteers (PDF, LIT, TeX and many others) is usually accepted, as long as they also supply an ASCII version where possible.

Initially, the books were mostly in English. As the original Project Gutenberg is based in the United States, its first focus was the English-speaking community in the country and worldwide. In October 1997, Michael Hart expressed his intention to digitize ebooks in other languages. In early 1998, the catalog had a few titles in French (10 titles), German, Italian, Spanish and Latin. In July 1999, Michael wrote: "I am publishing in one new language per month right now, and will continue as long as possible."

In the 2000s, multilingualism became a priority for Project Gutenberg, like internationalization, with Project Gutenberg Australia (created in August 2001), Project Gutenberg Europe (created in January 2004), Project Gutenberg Canada (created in July 2007), and others to come.

The launching of Project Gutenberg Europe and Distributed Proofreaders Europe (DP Europe) by Project Rastko was an important step. Founded in 1997, Project Rastko is a non-governmental cultural and educational project. One of its goals is the online publishing of Serbian culture. It is part of the Balkans Cultural Network Initiative, a regional cultural network for the Balkan peninsula in south-eastern Europe.

DP Europe has used the software of the original Distributed Proofreaders, launched in 2000 to share proofreading among a number of volunteers. Since the beginning, DP Europe has been a multilingual website, with its main pages translated into several European languages by volunteer translators. In April 2004, DP Europe was available in 12 languages. The long-term goal was 60 languages and 60 linguistic teams in the main European languages. DP Europe supports Unicode instead of ASCII, to be able to proofread ebooks in numerous languages.

First published in January 1991, Unicode "provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language" (excerpt from the website). This double-byte platform-independent encoding provides a basis for the processing, storage and interchange of text data in any language, and any modern software and information technology protocols. Unicode is maintained by the Unicode Consortium, and is a component of the W3C (World Wide Web Consortium) specifications. In 2008, 50% of available documents on the internet were encoded in Unicode, with the other 50% encoded in ASCII.

In the original Project Gutenberg in the U.S., there were ebooks in 25

languages in February 2004, in 42 languages in July 2005, including Sanskrit and the Mayan languages, and in 50 languages in December 2006. The ten top languages were English, French, German, Finnish, Dutch, Spanish, Chinese, Italian, Portuguese and Tagalog.

[Many thanks to Russon Wooldridge and Mike Cook for revising previous versions of this section.]

FIRST MULTILINGUAL PROJECTS

= [Quote]

Tyler Chambers, who created the Human-Languages Page and the Internet Dictionary Project, wrote in September 1998: "Online, my work has been with making language information available to more people through a couple of my web-based projects. While I'm not multilingual, nor even bilingual, myself, I see an importance to language and multilingualism that I see in very few other areas. The internet has allowed me to reach millions of people and help them find what they're looking for, something I'm glad to do. (...) Overall, I think that the web has been great for language awareness and cultural issues — where else can you randomly browse for 20 minutes and run across three or more different languages with information you might potentially want to know?"

= Travlang

Travlang is a website dedicated to both travel and languages, created in 1994 by Michael C. Martin on his university's website when he was a student in physics. Travlang included one section called Foreign Languages for Travelers, with links to online tools to learn 60 languages. Another section, Translating Dictionaries, gave access to free dictionaries in a number of languages (Afrikaans, Czech, Danish, Dutch, Esperanto, Finnish, French, Frisian, German, Hungarian, Italian, Latin, Norwegian, Portuguese, Spanish). Other sections offered links to language dictionaries, translation services, language schools, and multilingual bookstores. In 1998, Travlang was still maintained by its founder, who had become a researcher in experimental physics at the Lawrence Berkeley National Laboratory, California.

Michael C. Martin wrote in August 1998: "I think the web is an ideal place to bring different cultures and people together, and that includes being multilingual. Our Travlang site is so popular because of this, and people desire to feel in touch with other parts of the world. (...) The internet is really a great tool for communicating with people you wouldn't have the opportunity to interact with otherwise. I truly enjoy the global collaboration that has made our Foreign Languages for Travelers pages possible." Regarding the internet and languages in general, "I think computerized full-text translations will become more common, enabling a lot of basic communications with even more people. This will also help bring the internet more completely to the non- English speaking world."

= The Human-Languages Page

Created by Tyler Chambers in May 1994, the Human-Languages Page (H-LP) was a comprehensive catalog of 1,800 language-related internet resources in 100 languages. In September 1998, there were six subject listings and two category listings. The six subject listings were: languages and literature, schools and institutions, linguistics resources, products and services, organizations, jobs and internships. The two category listings were: dictionaries, and language lessons.

Tyler Chambers' other language-related project was the Internet Dictionary Project (IDP), launched in 1995. As explained on the project's website in September 1998: "The Internet Dictionary Project's goal is to create royalty-free translating dictionaries through the help of the internet's citizens. This site allows individuals from all over the world to visit and assist in the translation of English words into other languages. The resulting lists of English words and their translated counterparts are then made available through this site to anyone, with no restrictions on their use. (...) The Internet Dictionary Project began in 1995 in an effort to provide a noticeably lacking resource to the internet community and to computing in general — free translating dictionaries. Not only is it helpful to the online community to have access to dictionary searches at their fingertips via the World Wide Web, it also sponsors the growth of computer software which can benefit from such dictionaries — from translating programs to spelling-checkers to language-education guides and more. By facilitating the creation of these dictionaries online by thousands of anonymous volunteers all over the internet, and by providing the results free-of-charge to anyone, the Internet Dictionary Project hopes to leave its mark on the internet and to inspire others to create projects which will benefit more than a corporation's gross

income."

Tyler wrote in an email interview in September 1998: "Multilingualism on the web was inevitable even before the medium 'took off', so to speak. 1994 was the year I was really introduced to the web, which was a little while after its christening but long before it was mainstream. That was also the year I began my first multilingual web project, and there was already a significant number of language-related resources online. This was back before Netscape even existed — Mosaic was almost the only web browser, and webpages were little more than hyperlinked text documents. As browsers and users mature, I don't think there will be any currently spoken language that won't have a niche on the web, from Native American languages to Middle Eastern dialects, as well as a plethora of 'dead' languages that will have a chance to find a new audience with scholars and others alike online. To my knowledge, there are very few language types which are not currently online: browsers currently have the capability to display Roman characters, Asian languages, the Cyrillic alphabet, Greek, Turkish, and more. Accent Software has a product called 'Internet with an Accent' which claims to be able to display over 30 different language encodings. If there are currently any barriers to any particular language being on the web, they won't last long. (...)

Online, my work has been with making language information available to more people through a couple of my web-based projects. While I'm not multilingual, nor even bilingual, myself, I see an importance to language and multilingualism that I see in very few other areas. The internet has allowed me to reach millions of people and help them find what they're looking for, something I'm glad to do. It has also made me somewhat of a celebrity, or at least a familiar name in certain circles — I just found out that one of my web projects had a short mention in Time Magazine's Asia and International issues. Overall, I think that the web has been great for language awareness and cultural issues — where else can you randomly browse for 20 minutes and run across three or more different languages with information you might potentially want to know? Communications mediums make the world smaller by bringing people closer together; I think that the web is the first (of mail, telegraph, telephone, radio, TV) to really cross national and cultural borders for the average person. Israel isn't thousands of miles away anymore, it's a few clicks away — our world may now be small enough to fit inside a computer screen."

How about the future? "I think that the future of the internet is even more multilingualism and cross-cultural exploration and understanding than we've already seen. But the internet will only be the medium by which this information is carried; like the paper on which a book is written, the internet itself adds very little to the content of information, but adds tremendously to its value in its ability to communicate that information. To say that the internet is spurring multilingualism is a bit of a misconception, in my opinion — it is communication that is spurring multilingualism and cross-cultural exchange, the internet is only the latest mode of communication which has made its way down to the (more-or-less) common person. The internet has a long way to go before being ubiquitous around the world, but it, or some related progeny, likely will. Language will become even more important than it already is when the entire planet can communicate with everyone else (via the web, chat, games, e-mail, and whatever future applications haven't even been invented yet), but I don't know if this will lead to stronger language ties, or a consolidation of languages until only a few, or even just one remain. One thing I think is certain is that the internet will forever be a record of our diversity, including language diversity, even if that diversity fades away. And that's one of the things I love about the internet — it's a global model of the saying 'it's not really gone as long as someone remembers it'. And people do remember."

In spring 2001, the Human-Languages Page merged with the Languages Catalog, a section of the WWW Virtual Library, to become iLoveLanguages, In September 2003, iLoveLanguages provided an index of 2,000 linguistic resources in 100 languages. As for the Internet Dictionary Project, Tyler ran out of time to manage this project, and removed the ability to update the dictionaries in January 2007. People can still search the available dictionaries or download the archived files.

= NetGlos

Launched in 1995 by the WorldWide Language Institute (WWLI), an institute providing language instruction via the web, NetGlos (which stands for: Multilingual Glossary of Internet Terminology) has been compiled as a voluntary, collaborative project by a number of translators and other language professionals. In September 1998, NetGlos was available in the following languages: Chinese, Croatian, English, Dutch/Flemish, French, German, Greek, Hebrew, Italian, Maori, Norwegian, Portuguese, and Spanish.

Brian King, director of the WorldWide Language Institute, wrote in September 1998 in an email interview: "Although English is still the most important language used on the web, and the internet in general, I believe that multilingualism is an inevitable part of the future direction of cyberspace. Here

are some of the important developments that I see as making a multilingual web become a reality:

1. <Popularization of information technology.> Computer technology has traditionally been the sole domain of a 'techie' elite, fluent in both complex programming languages and in English — the universal language of science and technology. Computers were never designed to handle writing systems that couldn't be translated into ASCII. There wasn't much room for anything other than the 26 letters of the English alphabet in a coding system that originally couldn't even recognize acute accents and umlauts — not to mention non-alphabetic systems like Chinese. But tradition has been turned upside down. Technology has been popularized. GUIs (graphical user interfaces) like Windows and Macintosh have hastened the process (and indeed it's no secret that it was Microsoft's marketing strategy to use their operating system to make computers easy to use for the average person). These days this ease of use has spread beyond the PC to the virtual, networked space of the internet, so that now non-programmers can even insert Java applets into their webpages without understanding a single line of code.

2. <Competition for a chunk of the 'global market' by major industry players.> An extension of (local) popularization is the export of information technology around the world. Popularization has now occurred on a global scale and English is no longer necessarily the lingua franca of the user. Perhaps there is no true lingua franca, but only the individual languages of the users. One thing is certain — it is no longer necessary to understand English to use a computer, nor it is necessary to have a degree in computer science. A pull from non- English-speaking computer users and a push from technology companies competing for global markets has made localization a fast growing area in software and hardware development. This development has not been as fast as it could have been. The first step was for ASCII to become Extended ASCII. This meant that computers could begin to start recognizing the accents and symbols used in variants of the English alphabet — mostly used by European languages. But only one language could be displayed on a page at a time.

3. <Technological developments.> The most recent development is Unicode. Although still evolving and only just being incorporated into the latest software, this new coding system translates each character into 16 bytes. Whereas 8-byte Extended ASCII could only handle a maximum of 256 characters, Unicode can handle over 65,000 unique characters and therefore potentially accommodate all of the world's writing systems on the computer. So now the tools are more or less in place. They are still not perfect, but at last we can at least surf the web in Chinese, Japanese, Korean, and numerous other languages that don't use the Western alphabet. As the internet spreads to parts of the world where English is rarely used — such as China, for example, it is natural that Chinese, and not English, will be the preferred choice for interacting with it. For the majority of the users in China, their mother tongue will be the only choice. There is a change-over period, of course. Much of the technical terminology on the web is still not translated into other languages. And as we found with our Multilingual Glossary of Internet Terminology — known as NetGlos — the translation of these terms is not always a simple process. Before a new term becomes accepted as the 'correct' one, there is a period of instability where a number of competing candidates are used. Often an English loan word becomes the starting point — and in many cases the endpoint. But eventually a winner emerges that becomes codified into published technical dictionaries as well as the everyday interactions of the nontechnical user. The latest version of NetGlos is the Russian one and it should be available in a couple of weeks or so [at the end of September 1998]. It will no doubt be an excellent example of the ongoing, dynamic process of 'russification' of web terminology.

4. <Linguistic democracy.> Whereas 'mother-tongue education' was deemed a human right for every child in the world by a UNESCO report in the early '50s, 'mother-tongue surfing' may very well be the Information Age equivalent. If the internet is to truly become the Global Network that it is promoted as being, then all users, regardless of language background, should have access to it. To keep the internet as the preserve of those who, by historical accident, practical necessity, or political privilege, happen to know English, is unfair to those who don't.

5. <Electronic commerce.> Although a multilingual web may be desirable on moral and ethical grounds, such high ideals are not enough to make it other than a reality on a small-scale. As well as the appropriate technology being available so that the non-English speaker can go, there is the impact of 'electronic commerce' as a major force that may make multilingualism the most natural path for cyberspace. Sellers of products and services in the virtual global marketplace into which the internet is developing must be prepared to deal with a virtual world that is just as multilingual as the physical world. If they want to be successful, they had better make sure they are speaking the languages of their customers!"

How about the future of the WorldWide Language Institute? "As a company that derives its very existence from the importance attached to languages, I believe the future will be an exciting and challenging one. But it will be impossible to be complacent about our successes and accomplishments.

Technology is already changing at a frenetic pace. Lifelong learning is a strategy that we all must use if we are to stay ahead and be competitive. This is a difficult enough task in an English-speaking environment. If we add in the complexities of interacting in a multilingual/multicultural cyberspace, then the task becomes even more demanding. As well as competition, there is also the necessity for cooperation — perhaps more so than ever before. The seeds of cooperation across the internet have certainly already been sown. Our NetGlos Project has depended on the goodwill of volunteer translators from Canada, U.S., Austria, Norway, Belgium, Israel, Portugal, Russia, Greece, Brazil, New Zealand and other countries. I think the hundreds of visitors we get coming to the NetGlos pages everyday is an excellent testimony to the success of these types of working relationships. I see the future depending even more on cooperative relationships — although not necessarily on a volunteer basis."

= Logos

Logos is a global translation company with headquarters in Modena, Italy. In 1997, Logos had 200 in-house translators in Modena and 2,500 free-lance translators worldwide, who processed around 200 texts per day. The company made a bold move, and decided to put on the web the linguistic tools used by its translators, for the internet community to freely use them as well. The linguistic tools were the Logos Dictionary, a multilingual dictionary with 7 billion words (in fall 1998); the Logos Wordtheque, a multilingual library with 300 billion words extracted from translated novels, technical manuals and other texts; the Logos Linguistic Resources, a database of 500 glossaries; and the Logos Universal Conjugator, a database for verbs in 17 languages.

When interviewed by Annie Kahn in December 1997 for the French daily Le Monde, Rodrigo Vergara, head of Logos, explained: "We wanted all our translators to have access to the same translation tools. So we made them available on the internet, and while we were at it we decided to make the site open to the public. This made us extremely popular, and also gave us a lot of exposure. This move has in fact attracted many customers, and also allowed us to widen our network of translators, thanks to contacts made in the wake of the initiative."

In the same article, "Les mots pour le dire" (The Words to Tell it), Annie Kahn wrote: "The Logos site is much more than a mere dictionary or a collection of links to other online dictionaries. The cornerstone is the document search program, which processes a corpus of literary texts available free of charge on the web. If you search for the definition or the translation of a word ('didactique' [didactic], for example), you get not only the answer sought, but also a quote from one of the literary works containing the word (in our case, an essay by Voltaire). All it takes is a click on the mouse to access the whole text or even to order the book, including in foreign translations, thanks to a partnership agreement with the famous online bookstore Amazon.com. However, if no text containing the required word is found, the program acts as a search engine, sending the user to other web sources containing this word. In the case of certain words, you can even hear the pronunciation. If there is no translation currently available, the system calls on the public to contribute. Everyone can make suggestions, after which Logos translators check the suggested translations they receive."

ONLINE LANGUAGE DICTIONARIES

= [Quote]

WordReference.com was created in 1999 by Michael Kellogg, who wrote on his project's website: "I started this site in 1999 in an effort to provide free online bilingual dictionaries and tools to the world for free on the internet. The site has grown gradually ever since to become one of the most-used online dictionaries, and the top online dictionary for its language pairs of English-Spanish, English-French, English-Italian, Spanish-French, and Spanish-Portuguese. Today, I am happy to continue working on improving the dictionaries, its tools and the language forums. I really do enjoy creating new features to make the site more and more useful."

= From print versions

The first online language dictionaries stemmed from print versions, with websites launched in the mid-1990s.

On the website "Merriam-Webster Online: The Language Center", Merriam- Webster, a main publisher of English-language dictionaries, gave free access to online resources stemming from its print publications. The online resources were: Webster Dictionary, Webster Thesaurus, Webster's Third (a

lexical landmark), Guide to International Business Communications, Vocabulary Builder (with interactive vocabulary quizzes), and the Barnhart Dictionary Companion (hot new words). The goal was also to help track down definitions, spellings, pronunciations, synonyms, vocabulary exercises, and other key facts about words and language.

The "Dictionnaire Francophone en Ligne" was the web version of the "Dictionnaire Universel Francophone", published by Hachette, a major French publisher, and the University Agency for Francophony (AUF: Agence Universitaire de la Francophonie, also known as AUPELF-UREF). The dictionary included not only standard French but also the French- language words and expressions used worldwide. French is the official language of 49 states, with a number of them in Africa, and is spoken by 500 million people worldwide. The Agency of French-speaking Countries (Agence de la Francophonie), which has included the AUF, was founded in 1970 as an instrument of multilateral cooperation at the international level. As a side remark, English and French are the only official and/or cultural languages that are widely spread on five continents.

= Directories of dictionaries

Directories of dictionaries have been useful too, such as "Dictionnaires Électroniques" (Electronic Dictionaries), an online catalog of electronic dictionaries maintained by the French Section of the Swiss Federal Administration's Central Linguistic Services (SLC-f: Section Française des Services Linguistiques Centraux). The catalog included five main sections: abbreviations and acronyms, monolingual dictionaries, bilingual dictionaries, multilingual dictionaries, and geographical information. The catalog could also be searched by keywords.

Marcel Grangier was the head of the French Section of Central Linguistic Services, which means he was in charge of organizing translation matters into French for the linguistic services of the Swiss government. He wrote in January 1999: "Our website was first conceived as an intranet service for translators in Switzerland, who often deal with the same kind of material as the Federal government's translators. Some parts of it are useful to any translators, wherever they are. The section "Dictionnaires Électroniques" is only one section of the website. Other sections deal with administration, law, the French language, and general information. The site also hosts the pages of the Conference of Translation Services of European States (COTSOES). (...) To work without the internet is simply impossible now. Apart from all the tools used (email, the electronic press, services for translators), the internet is for us a vital and endless source of information in what I'd call the 'non-structured sector' of the web. For example, when the answer to a translation problem can't be found on websites presenting information in an organized way, in most cases search engines allow us to find the missing link somewhere on the network."

How about the future? "We can see multilingualism on the internet as a happy and irreversible inevitability. So we have to laugh at the doomsayers who only complain about the supremacy of English. Such supremacy isn't wrong in itself, because it is mainly based on statistics (more PCs per inhabitant, more people speaking English, etc.). The answer isn't to 'fight English', much less whine about it, but to build more sites in other languages. As a translation service, we also recommend that websites be multilingual. (...) The increasing number of languages on the internet is inevitable and can only boost multicultural exchanges. For this to happen in the best possible circumstances, we still need to develop tools to improve compatibility. Fully coping with accents and other characters is only one example of what can be done."

The section "Dictionnaires Électroniques" was later transfered on the website of the Conference of Translation Services of European States (COTSOES), when COTSOES launched its own website.

= The yourDictionary.com portal

Robert Beard, a language teacher at Bucknell University, in Lewisburg, Pennsylvania, created the website "A Web of Online Dictionaries" (WOD) in 1995. In September 1998, the website provided an index of 800 online dictionaries in 150 languages, as well as specific sections: multilingual dictionaries, specialized English dictionaries, thesauri and other vocabulary aids, language identifiers and guessers, an index of dictionary indices, the Web of Online Grammars, and the Web of Linguistic Fun (i.e. linguistics for non-specialists).

Robert Beard wrote in September 1998: "There was an initial fear that the web posed a threat to multilingualism on the web, since HTML and other programming languages are based on English and since there are simply more websites in English than any other language. However, my websites indicate that multilingualism is very much alive and the web may, in fact, serve as a vehicle for preserving many endangered languages. I now have links to dictionaries in 150 languages and

grammars of 65 languages. Moreover, the new attention paid by browser developers to the different languages of the world will encourage even more websites in different languages."

A few months later, Robert Beard co-founded a larger project, yourDictionary.com, that included his previous website and was launched in February 2000. He wrote in January 2000: "The new website is an index of 1,200+ dictionaries in more than 200 languages. Besides the WOD, the new website includes a word-of-the-day-feature, word games, a language chat room, the old 'Web of Online Grammars' (now expanded to include additional language resources), the 'Web of Linguistic Fun', multilingual dictionaries; specialized English dictionaries; thesauri and other vocabulary aids; language identifiers and guessers, and other features; dictionary indices. yourDictionary.com will hopefully be the premiere language portal and the largest language resource site on the web. It is now actively acquiring dictionaries and grammars of all languages with a particular focus on endangered languages. It is overseen by a blue ribbon panel of linguistic experts from all over the world. (...) Indeed, yourDictionary.com has lots of new ideas. We plan to work with the Endangered Language Fund in the U.S. and Britain to raise money for the Foundation's work and publish the results on our site. We will have language chatrooms and bulletin boards. There will be language games designed to entertain and teach fundamentals of linguistics. The Linguistic Fun page will become an online journal for short, interesting, yes, even entertaining, pieces on language that are based on sound linguistics by experts from all over the world."

How about the future of the web? "The web will be an encyclopedia of the world by the world for the world. There will be no information or knowledge that anyone needs that will not be available. The major hindrance to international and interpersonal understanding, personal and institutional enhancement, will be removed. It would take a wilder imagination than mine to predict the effect of this development on the nature of humankind."

= Terminological databases

Some terminological databases are run by international organizations in their own field of expertise, with free online versions, for example ILOTERM maintained by the International Labor Organization (ILO), TERMITE (ITU Telecommunication Terminology Database) maintained by the International Telecommunication Union (ITU), WHOTERM (WHO Terminology Information System) maintained by the World Health Organization (WHO), and Eurodicautom maintained by the European Commission.

ILOTERM is a quadrilingual (English, French, German, Spanish) terminology database maintained by the Terminology and Reference Unit of the Official Documentation Branch (OFFDOC) at the International Labor Office (ILO) in Geneva, Switzerland. As explained on its website, ILOTERM's primary purpose is to provide solutions, reflecting current usage, to terminological problems in the social and labor fields. Terms are entered in English with their French, Spanish and German equivalents. The database also includes records for the ILO structure and programs, official names of international institutions, national bodies and employers' and workers' organizations, and titles of international meetings.

TERMITE (which stands for: Telecommunication Terminology Database) is maintained by the Terminology, References and Computer Aids to Translation Section of the Conference Department at the International Telecommunication Union (ITU) in Geneva, Switzerland. It is a quadrilingual (English, French, Spanish, Russian) terminological database built on the content of all ITU printed glossaries since 1980, and updated with recent entries.

WHOTERM (which stands for: WHO Terminology Information System) is maintained by the World Health Organization (WHO) in Geneva, Switzerland. It has included: (a) the WHO General Dictionary Index (in English, with the French and Spanish equivalents); (b) three glossaries in English: Health for All, Programme Development and Management, and Health Promotion; (c) the WHO TermWatch, an awareness service from the Technical Terminology, reflecting the current WHO usage, but not necessarily terms officially approved by WHO, and links to health- related terminology.

Eurodicautom, a multilingual terminological database maintained by the Translation Service of the European Commission, was initially developed to assist in-house translators. The free online version was used by European Union officials and by language professionals throughout the world. Its contents were available in the eleven official languages of the European Union (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish), plus Latin. Eurodicautom covered "a broad spectrum of human knowledge", mainly relating to economy, science, technology and legislation in the European Union. In late 2003, the website announced the inclusion of the existing database into a larger terminological database that would also include databases from other official European institutions. The new terminological database would be available in more than 20 languages, because a number of Eastern European countries were expected to join the European Union in the near future,

thus the need for more languages than the eleven original ones. The European Union went from 15 country members to 25 country members in May 2004, and 27 country members in January 2007. The website of IATE (Inter-Active Terminology for Europe) was launched in March 2007 as an eagerly awaited free service on the web, with 1.4 million entries in 24 languages.

= Wikipedia

Wikipedia was launched in January 2001 by Jimmy Wales and Larry Sanger (Larry resigned later on). It has quickly grown into the largest reference website on the internet, financed by donations, with no advertising. Its multilingual content is free and written collaboratively by people worldwide, who contribute under a pseudonym. Its website is a wiki, which means that anyone can edit, correct and improve information throughout the encyclopedia. The articles stay the property of their authors, and can be freely used according to the GFDL (GNU Free Documentation License).

Wikipedia had 1.3 million articles (by 13,000 contributors) in 100 languages in December 2004, 6 million articles in 250 languages in December 2006, and 7 million articles in 192 languages in May 2007, including 1.8 million articles in English, 589,000 articles in German, 500,000 articles in French, 260,000 articles in Portuguese, and 236,000 articles in Spanish. In August 2009, Wikipedia was among the top five websites in the world, with a total of 330 million visitors a month.

Wikipedia is hosted by the Wikimedia Foundation, founded in June 2003, which has run a number of other projects, beginning with Wiktionary (launched in December 2002) and Wikibooks (launched in June 2003), followed by Wikiquote, Wikisource (texts from public domain), Wikimedia Commons (multimedia), Wikispecies (animals and plants), Wikinews, Wikiversity (textbooks), and Wiki Search (search engine).

LEARNING LANGUAGES ONLINE

= [Quote]

Robert Beard, a language teacher at Bucknell University, in Lewisburg, Pennsylvania, wrote in September 1998: "As a language teacher, the web represents a plethora of new resources produced by the target culture, new tools for delivering lessons (interactive Java and Shockwave exercises) and testing, which are available to students any time they have the time or interest — 24 hours a day, 7 days a week. It is also an almost limitless publication outlet for my colleagues and I, not to mention my institution. (…) Ultimately all course materials, including lecture notes, exercises, moot and credit testing, grading, and interactive exercises will be far more effective in conveying concepts that we have not even dreamed of yet."

= CTI Centre for Modern Languages

Since its inception in 1989, the CTI (Computer in Teaching Initiative) Centre for Modern Languages, based in the Language Institute at the University of Hull, United Kingdom, aims to promote and encourage the use of computers in language learning and teaching. The CTI Centre provides information on how computer-assisted language learning (CALL) can be effectively integrated into existing courses. It offers support to language lecturers who are using computers in their teaching, or who wish to use them.

June Thompson, manager of the CTI Centre, wrote in December 1998: "The internet has the potential to increase the use of foreign languages, and our organization certainly opposed any trend towards the dominance of English as the language of the internet. The use of the internet has brought an enormous new dimension to our work of supporting language teachers in their use of technology in teaching."

How about the future? "I suspect that for some time to come, the use of internet-related activities for languages will continue to develop alongside other technology-related activities (e.g. use of CD-ROMs — not all institutions have enough networked hardware). In the future I can envisage use of internet playing a much larger part, but only if such activities are pedagogy-driven. Our organization is closely associated with the WELL project which devotes itself to these issues."

The WELL (Web Enhanced Language Learning) project was a project from EUROCALL (European Association for Computer-Assisted Language Learning). It ran from 1997 to 2000 in the United Kingdom to provide access to high-quality web resources in 12 languages. The resources were selected and described by subject experts, with information and examples on how to use them for teaching and learning.

More generally, EUROCALL's goal is to promote the use of foreign languages within Europe, to provide a European focus for all aspects of the use of technology for language learning, and to enhance the quality, dissemination and efficiency of CALL materials. Another project of EUROCALL is CAPITAL (Computer-Assisted Pronunciation Investigation Teaching and Learning), run by a group of researchers and practitioners interested in using computers in this field.

= LINGUIST List

The LINGUIST List was founded by Anthony Rodriques Aristar in 1990 at the University of Western Australia, with 60 subscribers, before moving from Australia to Texas A&M University in 1991. In 1997, emails sent to the distribution list were also available on the list's own website, in the following sections: the profession (conferences, linguistic associations, programs), research and research support (papers, dissertation abstracts, projects, bibliographies, topics, texts), publications, pedagogy, language resources (languages, language families, dictionaries, regional information), and computer support (fonts and software). The LINGUIST List is a component of the WWW Virtual Library for linguistics.

Helen Dry, moderator of the LINGUIST List, wrote in August 1998: "The LINGUIST List, which I moderate, has a policy of posting in any language, since it's a list for linguists. However, we discourage posting the same message in several languages, simply because of the burden extra messages put on our editorial staff. (We are not a bounce- back list, but a moderated one. So each message is organized into an issue with like messages by our student editors before it is posted.) Our experience has been that almost everyone chooses to post in English. But we do link to a translation facility that will present our pages in any of five languages; so a subscriber need not read LINGUIST in English unless s/he wishes to. We also try to have at least one student editor who is genuinely multilingual, so that readers can correspond with us in languages other than English."

She added in July 1999: "We are beginning to collect some primary data. For example, we have searchable databases of dissertation abstracts relevant to linguistics, of information on graduate and undergraduate linguistics programs, and of professional information about individual linguists. The dissertation abstracts collection is, to my knowledge, the only freely available electronic compilation in existence."

MINORITY LANGUAGES ON THE WEB

= [Quote]

Caoimhín Ó Donnaíle has taught computing — through the Gaelic language — at the Institute Sabhal Mór Ostaig, on the Island of Skye, in Scotland. He has also maintained the bilingual (English, Gaelic) college website, which is the main site worldwide with information on Scottish Gaelic. He wrote in May 2001: "Students do everything by computer, use Gaelic spell-checking, a Gaelic online terminology database. There are more hits on our website. There is more use of sound. Gaelic radio (both Scottish and Irish) is now available continuously worldwide via the internet. A major project has been the translation of the Opera web browser into Gaelic — the first software of this size available in Gaelic."

= The Ethnologue

Published by SIL International (SIL was initially known as the Summer Institute of Linguistics), "The Ethnologue: Languages of the World" is an encyclopedic reference work cataloging all of the world’s 6,909 known living languages. The 16th edition was published in 2009, in print and on the web. The Ethnologue has been an active research project for more than fifty years. Thousands of linguists have contributed to the Ethnologue worldwide. A new edition is published approximately every four years.

The Ethnologue was founded in 1951 by Richard Pittman, who was motivated by the desire to share information on language development needs around the world with his colleagues at SIL International as well as with other language researchers. Richard Pittman was the editor of the 1st to 7th editions (1951-1969).

Barbara Grimes was the editor of the 8th to 14th editions (1971-2000). She wrote in January 2000: "It is a catalog of the languages of the world, with information about where they are spoken, an estimate of the number of speakers, what language family they are in, alternate names, names of dialects, other socio-linguistic and demographic information, dates of published Bibles, a name index, a language family index, and language maps." In 1971, information was expanded from primarily minority

languages to encompass all known languages of the world. Between 1967 and 1973, she completed an in-depth revision of the information on Africa, the Americas, the Pacific, and a few countries of Asia. During her years as editor, the number of identified languages grew from 4,493 to 6,809. The information recorded on each language expanded so that the published work more than tripled in size.

In 2000, Raymond Gordon Jr. became the third editor of the Ethnologue and produced the 15th edition (2005). Shortly after the publication of the 15th edition, Paul Lewis became the editor, responsible for general oversight and research policy. He installed Conrad Hurd as managing editor, responsible for operations and database management, and Raymond Gordon as senior research editor, leading a team of regional and language-family focused research editors.

In the Introduction of its latest edition (16th edition, 2009), the Ethnologue defines a language as such: "How one chooses to define a language depends on the purposes one has in identifying that language as distinct from another. Some base their definition on purely linguistic grounds. Others recognize that social, cultural, or political factors must also be taken into account. In addition, speakers themselves often have their own perspectives on what makes a particular language uniquely theirs. Those are frequently related to issues of heritage and identity much more than to the linguistic features of the language(s) in question."

As explained in the Introduction, one feature of the database since its inception has been a system of three-letter language identifiers, that appeared in the publication itself from the 10th edition (1984) onwards. "In 1998, the International Organization for Standardization (ISO) adopted ISO 639-2, a standard for three-letter language identifiers. The standard is based on a convergence of ISO 639-1 (an earlier standard for two-letter language identifiers adopted in 1988) and of ANSI Z39.53 (also known as the MARC language codes, a set of three-letter identifiers developed within the library community and adopted as an American National Standard in 1987). The ISO 639-2 standard was insufficient for many purposes since it has identifiers for fewer than 400 individual languages. Thus in 2002, ISO TC37/SC2 formally invited SIL International to prepare a new standard that would reconcile the complete set of codes used in the Ethnologue with the codes already in use in the earlier ISO standard. In addition, codes developed by Linguist List to handle ancient and constructed languages were to be incorporated. The result, which was officially approved by the subscribing national standards bodies in 2006 and published in 2007, is a standard named ISO 639-3 that provides standardized three- letter codes for identifying nearly 7,500 languages (ISO 2007). SIL International was named as the registration authority for the ISO 639-3 standard inventory of language identifiers and administers the annual cycle for changes and updates. This edition of Ethnologue is the second to use the ISO 639-3 language identifiers. In the fifteenth edition they had the status of Draft International Standard. In this edition they are based on the standard as originally adopted plus the 2006 series of adopted change requests (released August 2007) and the 2007 series of adopted change requests (released January 2008). Information about the ISO 639-3 standard and procedures for requesting additions, deletions, and other modifications to the ISO 639-3 inventory of identified languages can be found at the ISO 639-3 website: http://www.sil.org/iso639-3."

= Experiences

Caoimhín Ó Donnaíle has taught computing - through the Gaelic language - at the Institute Sabhal Mór Ostaig, on the Island of Skye, in Scotland. He has also maintained the bilingual (English, Gaelic) college website, which is the main site worldwide with information on Scottish Gaelic, as well as the bilingual webpage European Minority Languages, a list of minority languages by alphabetic order and by language family. He wrote in May 2001: "There has been a great expansion in the use of information technology in our college. Far more computers, more computing staff, flat screens. Students do everything by computer, use Gaelic spell-checking, and a Gaelic online terminology database. There are more hits on our website. There is more use of sound. Gaelic radio (both Scottish and Irish) is now available continuously worldwide via the internet. A major project has been the translation of the Opera web browser into Gaelic — the first software of this size available in Gaelic."

How about the internet and endangered languages? "I would emphasize the point that as regards the future of endangered languages, the internet speeds everything up. If people don't care about preserving languages, the internet and accompanying globalization will greatly speed their demise. If people do care about preserving them, the internet will be a tremendous help."

Guy Antoine is the founder of Windows on Haiti, a reference website about Haitian culture. He wrote in November 1999: "In Windows on Haiti, the primary language of the site is English, but one will equally find a center of lively discussion conducted in 'Kreyòl'. In addition, one will find documents related to Haiti in French, in the old colonial Creole, and I am open to publishing others in Spanish and other languages. I do not offer any sort of translation, but multilingualism is alive and well at the site, and I predict that this will increasingly become the norm throughout the web."

Guy added in June 2001: "Kreyòl is the only national language of Haiti, and one of its two official languages, the other being French. It is hardly a minority language in the Caribbean context, since it is spoken by eight to ten million people. (...) I have taken the promotion of Kreyòl as a personal cause, since that language is the strongest of bonds uniting all Haitians, in spite of a small but disproportionately influential Haitian elite's disdainful attitude to adopting standards for the writing of Kreyòl and supporting the publication of books and official communications in that language. For instance, there was recently a two-week book event in Haiti's Capital and it was promoted as 'Livres en Folie' ('A mad feast for books'). Some 500 books from Haitian authors were on display, among which one could find perhaps 20 written in Kreyòl. This is within the context of France's major push to celebrate Francophony among its former colonies. This plays rather well in Haiti, but directly at the expense of Creolophony. What I have created in response to those attitudes are two discussion forums on my website, Windows on Haiti, held exclusively in Kreyòl. One is for general discussions on just about everything but obviously more focused on Haiti's current socio-political problems. The other is reserved only to debates of writing standards for Kreyòl. Those debates have been quite spirited and have met with the participation of a number of linguistic experts. The uniqueness of these forums is their non- academic nature."

Robert Beard, co-founder of the yourDictionary.com portal, wrote in January 2000: "While English still dominates the web, the growth of monolingual non-English websites is gaining strength with the various solutions to the font problems. Languages that are endangered are primarily languages without writing systems at all (only 1/3 of the world's 6,000+ languages have writing systems). I still do not see the web contributing to the loss of language identity and still suspect it may, in the long run, contribute to strengthening it. More and more Native Americans, for example, are contacting linguists, asking them to write grammars of their language and help them put up dictionaries. For these people, the web is an affordable boon for cultural expression."

LOCALIZATION AND INTERNATIONALIZATION

= [Quote]

Peter Raggett, deputy-head (and then head) of the Central Library at the OECD (Organization for Economic Cooperation and Development), wrote in August 1999: "I think it is incumbent on European organizations and businesses to try and offer websites in three or four languages if resources permit. In this age of globalization and electronic commerce, businesses are finding that they are doing business across many countries. Allowing French, German, Japanese speakers to easily read one's website as well as English speakers will give a business a competitive edge in the domain of electronic trading."

= [Text]

In 1999, the subtitle of Babel's website was: "Towards communicating on the internet in any language..." Babel was a joint project from Alis Technologies and the Internet Society to contribute to the internationalization of the internet. Babel offered a multilingual website (English, French, German, Italian, Portuguese, Spanish and Swedish), with information about the world's languages, and a typographical and linguistic glossary. "The Internet and Multilingualism" section gave information on how to develop a multilingual website, and how to code the "world's writing".

Bill Dunlap, founder of Euro-Marketing Associates, a company based in San Francisco and Paris, launched the international marketing consultancy Global Reach as a methodology for U.S. companies to expand their internet presence into an international framework. This included translating a website into other languages, actively promoting it, and using local online banner advertising to increase local website traffic.

Bill Dunlap explained in December 1998: "Promoting your website is at least as important as creating it, if not more important. You should be prepared to spend at least as much time and money in promoting your website as you did in creating it in the first place. With the Global Reach program, you can have it promoted in countries where English is not spoken, and achieve a wider audience... and more sales. There are many good reasons for taking the online international market seriously. Global Reach is a means for you to extend your website to many countries, speak to online visitors in their own language and reach online markets there. (...)

Since 1981, when my professional life started, I've been involved with bringing American companies in Europe. This is very much an issue of language, since the products and their marketing have to be in the languages of Europe in order for them to be visible here. Since the web became popular in 1995 or

so, I've turned these activities to their online dimension, and have come to champion European e-commerce among my fellow American compatriots. Most lately at Internet World in New York, I spoke about European e-commerce and how to use a website to address the various markets in Europe."

Bill added in July 1999: "After a website's home page is available in several languages, the next step is the development of content in each language. A webmaster will notice which languages draw more visitors (and sales) than others, and these are the places to start in a multilingual web promotion campaign. At the same time, it is always good to increase the number of languages available on a website: just a home page translated into other languages would do for a start, before it becomes obvious that more should be done to develop a certain language branch on a website."

The World Wide Web Consortium (W3C) was founded in October 1994 to develop interoperable technologies (specifications, guidelines, software, and tools) for the web, for example specifications for markup languages (HTML, XML, and others), and to act as a forum for information, commerce, communication and collective understanding. In 1998, the section Internationalization/Localization gave a definition of protocols used for internationalization/localization: HTML, base character set, new tags and attributes, HTTP, language negotiation, URLs & other identifiers including non-ASCII characters, etc. It also offered some help with creating a multilingual website.

The Localisation Industry Standards Association (LISA) was created in the mid-1990s as a forum for "software publishers, hardware manufacturers, localization service vendors, and an increasing number of companies from related IT sectors." LISA has defined its mission as "promoting the localization and internationalization industry and providing a mechanism and services to enable companies to exchange and share information on the development of processes, tools, technologies and business models connected with localization, internationalization and related topics". Its website was first housed and maintained by the University of Geneva, Switzerland.

Launched in January 1999 by the European Commission, the website HLTCentral (HLT: Human Language Technologies) gave a short definition of language engineering: "Through language engineering we can find ways of living comfortably with technology. Our knowledge of language can be used to develop systems that recognize speech and writing, understand text well enough to select information, translate between different languages, and generate speech as well as the printed world. By applying such technologies we have the ability to extend the current limits of our use of language. Language enabled products will become an essential and integral part of everyday life."

MACHINE TRANSLATION

= [Quote]

Tim McKenna is an author who thinks and writes about the complexity of truth in a world of flux. He wrote in October 2000: "When software gets good enough for people to chat or talk on the web in real time in different languages, then we will see a whole new world appear before us. Scientists, political activists, businesses and many more groups will be able to communicate immediately without having to go through mediators or translators."

= A definition

Machine translation can be defined as the automated process of translating a text from one language to another language. MT analyzes the text in the source language and automatically generates the corresponding text in the target language. With the lack of any human intervention during the translation process, machine translation (MT) differs from computer-assisted translation (CAT), which involves some interaction between the translator and the computer.

As explained on the website of SYSTRAN, a company specializing in translation software, "machine translation software translates one natural language into another natural language. MT takes into account the grammatical structure of each language and uses rules to transfer the grammatical structure of the source language (text to be translated) into the target language (translated text). MT cannot replace a human translator, nor is it intended to."

The website of the European Association for Machine Translation (EAMT) gives the following definition: "Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. One of the very earliest pursuits in computer science, MT has proved to be an elusive goal, but today a number of systems are available which produce output which, if not perfect, is of sufficient quality to be useful for certain specific applications, usually in the domain

of technical documentation. In addition, translation software packages which are designed primarily to assist the human translator in the production of translations are enjoying increasing popularity within professional translation organizations."

Machine translation is the earliest type of natural language processing, as stated on the website of Globalink, a company offering language translation software and services: "From the very beginning, machine translation (MT) and natural language processing (NLP) have gone hand-in-hand with the evolution of modern computational technology. The development of the first general-purpose programmable computers during World War II was driven and accelerated by Allied cryptographic efforts to crack the German Enigma machine and other wartime codes. Following the war, the translation and analysis of natural language text provided a testbed for the newly emerging field of Information Theory.

During the 1950s, research on Automatic Translation (known today as Machine Translation, or 'MT') took form in the sense of literal translation, more commonly known as word-for-word translations, without the use of any linguistic rules. The Russian project initiated at Georgetown University in the early 1950s represented the first systematic attempt to create a demonstrable machine translation system. Throughout the decade and into the 1960s, a number of similar university and government-funded research efforts took place in the United States and Europe. At the same time, rapid developments in the field of Theoretical Linguistics, culminating in the publication of Noam Chomsky's "Aspects of the Theory of Syntax" (1965), revolutionized the framework for the discussion and understanding of the phonology, morphology, syntax and semantics of human language.

In 1966, the U.S. government-issued ALPAC (Automatic Language Processing Advisory Committee) report offered a prematurely negative assessment of the value and prospects of practical machine translation systems, effectively putting an end to funding and experimentation in the field for the next decade. It was not until the late 1970s, with the growth of computing and language technology, that serious efforts began once again. This period of renewed interest also saw the development of the Transfer model of machine translation and the emergence of the first commercial MT systems. While commercial ventures such as SYSTRAN and METAL began to demonstrate the viability, utility and demand for machine translation, these mainframe-bound systems also illustrated many of the problems in bringing MT products and services to market. High development cost, labor-intensive lexicography and linguistic implementation, slow progress in developing new language pairs, inaccessibility to the average user, and inability to scale easily to new platforms are all characteristics of these second-generation systems."

As explained in August 1998 by Eduard Hovy, head of the Natural Language Group at USC/ISI (University of Southern California/Information Sciences Institute), machine translation implies "language-related applications/functionalities that are not translation, such as information retrieval (IR) and automated text summarization (SUM). You would not be able to find anything on the Web without IR! — all the search engines (AltaVista, Yahoo!, etc.) are built upon IR technology. Similarly, though much newer, it is likely that many people will soon be using automated summarizers to condense (or at least, to extract the major contents of) single (long) documents or lots of (any length) ones together."

= Experiences

In December 1997, AltaVista, a leading search engine, was the first to launch a free translation software with Babel Fish — also called AltaVista Translation —, which could translate webpages (up to three pages at the same time) from English into French, German, Italian, Portuguese or Spanish, and vice versa. The software was developed by SYSTRAN (an acronym for System Translation), a company specializing in machine translation software. SYSTRAN's headquarters are located in Soisy-sous-Montmorency, near Paris, France. Sales, marketing, and research and development are based in its subsidiary in La Jolla, California.

This initiative was followed by other translation software developed by Alis Technologies, Globalink, Lernout & Hauspie, and Softissimo, with free and/or paid versions on the web.

Based in Montreal, Quebec, Alis Technologies has specialized in development and marketing of language handling solutions and services, particularly language implementation in the information technology industry. Alis Translation Solutions (ATS) has offered applications in a number of languages, and tools and services to improve the quality of translations. Language Technology Solutions (LTS) has marketed advanced tools and services for language engineering and information technology (90 languages covered).

Based in Ieper, Belgium, and Burlington, Massachusetts, Lernout & Hauspie (L&H) was a leader in advanced speech technology for commercial applications and products, with four core technologies:

automatic speech recognition (ASR), text-to-speech (TTS), text-to-text (TTT), and digital speech compression (DSC). Its ASR, TTS and DSC technologies were licensed to companies in telecommunications, computers and multimedia, consumer electronics and automotive electronics. Its TTT translation services were provided to IT companies, and vertical and automation markets. The Machine Translation Group created by Lernout & Hauspie included L&H Language Technology, AppTek, AILogic, NeocorTech, and Globalink. Lernout & Hauspie was later bought by Nuance Communications.

Globalink, a company created in 1990 in the U.S., focused on language translation software and services, i.e. customized translation solutions built around software products, online options, and professional translation services. The software products were available in Spanish, French, Portuguese, German, Italian and English, for individuals, small businesses, multinational corporations and governments, from a stand-alone product giving a fast draft translation to a full system managing professional translations.

As explained on the company website in 1998, "with Globalink's translation applications, the computer uses three sets of data: the input text, the translation program and permanent knowledge sources (containing a dictionary of words and phrases of the source language), and information about the concepts evoked by the dictionary and rules for sentence development. These rules are in the form of linguistic rules for syntax and grammar, and some are algorithms governing verb conjugation, syntax adjustment, gender and number agreement and word re-ordering. Once the user has selected the text and set the machine translation process in motion, the program begins to match words of the input text with those stored in its dictionary. Once a match is found, the application brings up a complete record that includes information on possible meanings of the word and its contextual relationship to other words that occur in the same sentence. The time required for the translation depends on the length of the text. A three-page, 750-word document takes about three minutes to render a first draft translation."

At the headquarters of the World Health Organization (WHO) in Geneva, Switzerland, the Computer-assisted Translation and Terminology Unit (CTT) has been a pioneer since 1997 in assessing technical options for using computer-assisted translation (CAT) systems based on translation memory (TM). With such systems, translators can access previous translations from portions of the text; accept, reject or modify them; and add the new translation to the memory, thus enriching it for future reference. By archiving the daily output, the translator helps in building an extensive translation memory and in solving a number of translation issues. Several projects have been under way at the CTT for electronic document archiving and retrieval, bilingual/multilingual text alignment, computer-assisted translation, translation memory and terminology database management, and speech recognition.

The Pan American Health Organization (PAHO) in Washington, D.C. has developed its own machine translation software, as a common work from its own computational linguists, translators, and system programmers. The PAHO Translation Unit has used SPANAM (Spanish to English) from 1980 and ENGSPAN (English to Spanish) from 1985, to process over 25 million words between 1980 and 1998. Staff translators and free-lance translators post-edit the raw output to produce high-quality translations with a 30-50% gain in productivity. The software is available in the LAN (Local Area Network) of PAHO Headquarters, and is regularly used by the staff of technical and administrative units. The software is also available in a number of PAHO field offices, and has been licensed to public and non-profit institutions in the U.S., Latin America, and Spain. The software was later renamed PAHOMTS, and has included new language pairs with Portuguese.

= Comments

Comments from ZDNN

In "Web Embraces Language Translation", an article published in ZDNN (ZDNetwork News) on 21 July 1998, Martha Stone explained: "Among the new products in the $10 billion language translation business are instant translators for websites, chat rooms, email and corporate intranets. The leading translation firms are mobilizing to seize the opportunities. Such as:

*SYSTRAN has partnered with AltaVista and reports between 500,000 and 600,000 visitors a day on babelfish.altavista.digital.com, and about 1 million translations per day — ranging from recipes to complete webpages. About 15,000 sites link to babelfish, which can translate to and from French, Italian, German, Spanish and Portuguese. The site plans to add Japanese soon. 'The popularity is simple. With the internet, now there is a way to use U.S. content. All of these contribute to this increasing demand,' said Dimitros Sabatakakis, group CEO of SYSTRAN, speaking from his Paris home.

*Alis technology powers the Los Angeles Times' soon-to-be launched language translation feature on

its site. Translations will be available in Spanish and French, and eventually, Japanese. At the click of a mouse, an entire webpage can be translated into the desired language.

*Globalink offers a variety of software and web translation possibilities, including a free email service and software to enable text in chat rooms to be translated.

But while these so-called 'machine' translations are gaining worldwide popularity, company execs admit they're not for every situation. Representatives from Globalink, Alis and SYSTRAN use such phrases as 'not perfect' and 'approximate' when describing the quality of translations, with the caveat that sentences submitted for translation should be simple, grammatically accurate and idiom-free. 'The progress on machine translation is moving at Moore's Law — every 18 months it's twice as good,' said Vin Crosbie, a web industry analyst in Greenwich, Conn. 'It's not perfect, but some [non-English speaking] people don't realize I'm using translation software.'

With these translations, syntax and word usage suffer, because dictionary-driven databases can't decipher between homonyms — for example, 'light' (as in the sun or light bulb) and 'light' (the opposite of heavy). Still, human translation would cost between $50 and $60 per webpage, or about 20 cents per word, SYSTRAN's Sabatakakis said. While this may be appropriate for static 'corporate information' pages, the machine translations are free on the web, and often less than $100 for software, depending on the number of translated languages and special features."

Comments from RALI

Despite the imminent outbreak of a universal translation machine announced at the end of the 1940s, machine translation hasn't produced good translations yet. Pierre Isabelle and Patrick Andries, two scientists from the RALI Laboratory (Laboratory for Applied Research in Computational Linguistics - Laboratoire de Recherche Appliquée en Linguistique Informatique) in Montreal, Quebec, explain the reasons for this failure in "La Traduction Automatique, 50 Ans Après" (Machine Translation, 50 Years Later), an article published in 1998 by Multimédium, a French-language online magazine: "The ultimate goal of building a machine capable of competing with a human translator remains elusive due to slow progress in research. (...) Recent research, based on large collections of texts called corpora — using either statistical or analogical methods — has promised to reduce the quantity of manual work required to build a machine translation (MT) system, but can't promise for sure a significant improvement in the quality of machine translation. (...) The use of MT will be more or less restricted to tasks of information assimilation or tasks of text distribution in restricted sub-languages."

According to Yehochua Bar-Hillel's ideas expressed in "The State of Machine Translation", an article published in 1951, Pierre Isabelle and Patrick Andries define three implementation strategies for machine translation: (a) a tool of information assimilation to scan multilingual data and supply rough translation, (b) situations of "restricted language" such as the METEO system which, since 1977, has translated the weather forecasts of the Canadian Ministry of Environment, (c) the human/machine coupling before, during and after the machine translation process, that may not save money if compared to traditional translation.

Pierre Isabelle and Patrick Andries favor "a workstation for the human translator" more than a "robot translator": "Recent research on the probabilist methods showed it was possible to modelize in an efficient way some simple aspects of the translation relationship between two texts. For example, methods were set up to calculate the correct alignment between the text sentences and their translation, that is, to identify the sentence(s) of the source text corresponding to each sentence of the translation. Applied on a large scale, these techniques can use the archives of a translation service to build a translation memory for recycling fragments from previous translations. Such systems are already available on the translation market (IBM Translation Manager II, Trados Translator's Workbench by Trados, RALI TransSearch, etc.) The latest research focuses on models that can automatically set up correspondences at a finer level than the sentence level, i.e. syntagms and words. The results let hope for a bunch of new tools for the human translator, including for the study of terminology, for dictation and translation typing, and for detectors of translation errors."

Comments from Randy Hobler

In September 1998, Randy Hobler was a consultant in internet marketing at Globalink, after working for IBM, Johnson & Johnson, Burroughs Wellcome, Pepsi, and Heublein. He wrote in an email interview: "We are rapidly reaching the point where highly accurate machine translation of text and speech will be so common as to be embedded in computer platforms, and even in chips in various ways. At that point, and as the growth of the web slows, the accuracy of language translation hits 98% plus, and the saturation of language pairs has covered the vast majority of the market, language transparency (any-language-to-any- language communication) will be too limiting a vision for those selling this technology. The next development will be 'transcultural, transnational transparency', in

which other aspects of human communication, commerce and transactions beyond language alone will come into play. For example, gesture has meaning, facial movement has meaning and this varies among societies. The thumb-index finger circle means 'OK' in the United States. In Argentina, it is an obscene gesture.

When the inevitable growth of multimedia, multilingual videoconferencing comes about, it will be necessary to 'visually edit' gestures on the fly. The MIT (Massachusetts Institute of Technology) Media Lab, Microsoft and many others are working on computer recognition of facial expressions, biometric access identification via the face, etc. It won't be any good for a U.S. business person to be making a great point in a web-based multilingual video conference to an Argentinian, having his words translated into perfect Argentinian Spanish if he makes the 'O' gesture at the same time. Computers can intercept this kind of thing and edit them on the fly.

There are thousands of ways in which cultures and countries differ, and most of these are computerizable to change as one goes from one culture to the other. They include laws, customs, business practices, ethics, currency conversions, clothing size differences, metric versus English system differences, etc. Enterprising companies will be capturing and programming these differences and selling products and services to help the peoples of the world communicate better. Once this kind of thing is widespread, it will truly contribute to international understanding."

= Machine translation R&D

Here is an overview of the work of four research centers, in Quebec (RALI Laboratory), California (Natural Language Group), Switzerland (ISSCO) and Japan (UNDL Foundation).

RALI Laboratory

In Montreal, Quebec, the RALI Laboratory (Laboratory of Applied Research in Computational Linguistics - Laboratoire de Recherche Appliquée en Linguistique Informatique) has worked in automatic text alignment, automatic text generation, automatic reaccentuation, language identification, and finite state transducers. RALI produces the "TransX family" of what it calls "a new generation" of translation support tools (TransType, TransTalk, TransCheck, and TransSearch), which are based on probabilistic translation models that automatically calculate correspondences between the text produced by a translator and the original text from the source language.

As explained on RALI's website in 1998: "(a) TransType speeds up the keying-in of a translation by anticipating a translator's choices and criticizing them when appropriate. In proposing its suggestions, TransType takes into account both the source text and the partial translation that the translator has already produced. (b) TransTalk is an automatic dictation system that makes use of a probabilistic translation model in order to improve the performance of its voice recognition model. (c) TransCheck automatically detects certain types of translation errors by verifying that the correspondences between the segments of a draft and the segments of the source text respect well- known properties of a good translation. (d) TransSearch allows translators to search databases of pre-existing translations in order to find ready-made solutions to all sorts of translation problems. In order to produce the required databases, the translations and the source language texts must first be aligned."

Natural Language Group

The Natural Language Group (NLG) at the Information Sciences Institute (ISI) of the University of Southern California (USC) has been involved in various aspects of computational/natural language processing: machine translation, automated text summarization, multilingual verb access and text management, development of large concept taxonomies (ontologies), discourse and text generation, construction of large lexicons for various languages, and multimedia communication.

Eduard Hovy, head of the Natural Language Group, explained in August 1998: "People will write their own language for several reasons — convenience, secrecy, and local applicability — but that does not mean that other people are not interested in reading what they have to say! This is especially true for companies involved in technology watch (say, a computer company that wants to know, daily, all the Japanese newspaper and other articles that pertain to what they make) or some Government Intelligence agencies (the people who provide the most up- to-date information for use by your government officials in making policy, etc.). One of the main problems faced by these kinds of people is the flood of information, so they tend to hire 'weak' bilinguals who can rapidly scan incoming text and throw out what is not relevant, giving the relevant stuff to professional translators. Obviously, a combination of SUM (automated text summarization) and MT (machine translation) will help here; since MT is slow, it helps if you can do SUM in the foreign language, and then just do a quick and dirty

MT on the result, allowing either a human or an automated IR-based text classifier to decide whether to keep or reject the article. For these kinds of reasons, the U.S. Government has over the past five years been funding research in MT, SUM, and IR (information retrieval), and is interested in starting a new program of research in Multilingual IR. This way you will be able to one day open Netscape or Explorer or the like, type in your query in (say) English, and have the engine return texts in *all* the languages of the world. You will have them clustered by subarea, summarized by cluster, and the foreign summaries translated, all the kinds of things that you would like to have."

Eduard Hovy added in August 1999: "Over the past 12 months I have been contacted by a surprising number of new information technology (IT) companies and startups. Most of them plan to offer some variant of electronic commerce (online shopping, bartering, information gathering, etc.). Given the rather poor performance of current non-research level natural language processing technology (when is the last time you actually easily and accurately found a correct answer to a question to the web, without having to spend too much time sifting through irrelevant information?), this is a bit surprising. But I think everyone feels that the new developments in automated text summarization, question analysis, and so on, are going to make a significant difference. I hope so!—but the level of performance is not available yet.

It seems to me that we will not get a big breakthrough, but we will get a somewhat acceptable level of performance, and then see slow but sure incremental improvement. The reason is that it is very hard to make your computer really 'understand' what you mean — this requires us to build into the computer a network of 'concepts' and their interrelationships that (at some level) mirror those in your own mind, at least in the subjects areas of interest. The surface (word) level is not adequate — when you type in 'capital of Switzerland', current systems have no way of knowing whether you mean 'capital city' or 'financial capital'. Yet the vast majority of people would choose the former reading, based on phrasing and on knowledge about what kinds of things one is likely to ask the web, and in what way. Several projects are now building, or proposing to build, such large 'concept' networks. This is not something one can do in two years, and not something that has a correct result. We have to develop both the network and the techniques for building it semi-automatically and self-adaptively. This is a big challenge."

Eduard Hovy added in September 2000: "I see a continued increase in small companies using language technology in one way or another: either to provide search, or translation, or reports, or some other communication function. The number of niches in which language technology can be applied continues to surprise me: from stock reports and updates to business-to-business communications to marketing…

With regard to research, the main breakthrough I see was led by a colleague at ISI (I am proud to say), Kevin Knight. A team of scientists and students last summer at Johns Hopkins University in Maryland developed a faster and otherwise improved version of a method originally developed (and kept proprietary) by IBM about 12 years ago. This method allows one to create a machine translation (MT) system automatically, as long as one gives it enough bilingual text. Essentially the method finds all correspondences in words and word positions across the two languages and then builds up large tables of rules for what gets translated to what, and how it is phrased.

Although the output quality is still low — no-one would consider this a final product, and no-one would use the translated output as is — the team built a (low-quality) Chinese-to-English MT system in 24 hours. That is a phenomenal feat — this has never been done before. (Of course, say the critics: you need something like 3 million sentence pairs, which you can only get from the parliaments of Canada, Hong Kong, or other bilingual countries; and of course, they say, the quality is low. But the fact is that more bilingual and semi-equivalent text is becoming available online every day, and the quality will keep improving to at least the current levels of MT engines built by hand. Of that I am certain.)

Other developments are less spectacular. There's a steady improvement in the performance of systems that can decide whether an ambiguous word such as "bat" means "flying mammal" or "sports tool" or "to hit"; there is solid work on cross-language information retrieval (which you will soon see in being able to find Chinese and French documents on the web even though you type in English-only queries), and there is some rather rapid development of systems that answer simple questions automatically (rather like the popular web system AskJeeves, but this time done by computers, not humans). These systems refer to a large collection of text to find 'factiods' (not opinions or causes or chains of events) in response to questions such as 'what is the capital of Uganda?' or 'how old is President Clinton?' or 'who invented the xerox process?', and they do so rather better than I had expected."

 # ISSCO

In Geneva, Switzerland, ISSCO (Dalle Molle Institute for Semantic and Cognitive Studies - Institut Dalle Molle pour les Études Sémantiques et Cognitives) is a research laboratory conducting basic and applied research in computational linguistics (CL) and artificial intelligence (AI), for a number of Swiss and European research projects. The University of Geneva has provided administrative support and infrastructure. Research is funded with grants and contracts with public and private bodies.

Created by the Foundation Dalle Molle in 1972 to conduct research in cognition and semantics, ISSCO has come to specialize in natural language processing, including multilingual language processing, in a number of areas: machine translation, linguistic environments, multilingual generation, discourse processing, data collection, etc. ISSCO is multi-disciplinary and multi-national. As explained on its website in 1998, "its staff and its visitors [are drawn] from the disciplines of computer science, linguistics, mathematics, psychology and philosophy. The long-term staff of the Institute is relatively small in number; with a much larger number of visitors coming for stays ranging from a month to two years. This ensures a continual exchange of ideas and encourages flexibility of approach amongst those associated with the Institute."

UNDL Foundation

The UNL (universal networking language) project was launched in the mid-1990s as a main digital metalanguage project by the Institute of Advanced Studies (IAS) of the United Nations University (UNU) in Tokyo, Japan. As explained on the bilingual (English, Japanese) website in 1998: "UNL is a language that — with its companion 'enconverter' and 'deconverter' software — enables communication among peoples of differing native languages. It will reside, as a plug-in for popular web browsers, on the internet, and will be compatible with standard network servers. The technology will be shared among the member states of the United Nations. Any person with access to the internet will be able to 'enconvert' text from any native language of a member state into UNL. Just as easily, any UNL text can be 'deconverted' from UNL into native languages. United Nations University's UNL Center will work with its partners to create and promote the UNL software, which will be compatible with popular network servers and computing platforms."

In 2000, 120 researchers worldwide were working on a multilingual project in 16 languages (Arabic, Brazilian, Chinese, English, French, German, Hindu, Indonesian, Italian, Japanese, Latvian, Mongolian, Russian, Spanish, Swahiki, and Thai). The UNDL Foundation (UNDL: Universal Networking Digital Language) was founded in January 2001 to develop and promote the UNL project.

CHRONOLOGY

[Each line begins with the year or the year/month.]

1968: ASCII is the first character set encoding.
1971: Project Gutenberg is the first digital library.
1974: The internet takes off.
1990: The web is invented by Tim Berners-Lee.
1991/01: Unicode is a universal character set encoding for all languages.
1993/11: Mosaic is the first web browser.
1994/05: The Human-Languages Page is a catalog of language-related internet resources.
1994/10: The World Wide Web Consortium will deal with internationalization and localization.
1994: Travland is dedicated to both travel and languages.
1995/12: The Kotoba Home Page deals with language issues using our keyboard.
1995: The Internet Dictionary Project works on creating free translating dictionaries.
1995: NetGlos is a multilingual glossary of internet terminology.
1995: Global Reach is a virtual consultancy stemming from Euro-Marketing Associates.
1995: LISA is the localization industry standards association.
1995: "The Ethnologue: Languages of the World" offers a free online version.
1996/04 : OneLook Dictionaries is a fast finder in online dictionaries.
1997/01: UNL (universal networking language) is a digital metalanguage project.
1997/12: AltaVista launches AltaVista Translation, also called Babel Fish.
1997: The Logos Dictionary goes online for free.
1999/12: Britannica.com is the first main English-language online encyclopedia.
1999/12: WebEncyclo is the first main French-language online encyclopedia.
1999: WordReference.com offers free online bilingual translating dictionaries.
2000/02: yourDictionary.com is a major language portal.
2000/07: Non-English-speaking internet users reach 50%.
2001/01: Wikipedia is a main free multilingual cooperative encyclopedia.

2001/01: The UNDL Foundation develops UNL, a digital metalanguage project.
2001/04: The Human-Languages Project becomes the iLoveLanguages portal.
2004/01: Project Gutenberg Europe is launched as a multilingual project.
2007/03: IATE is the new terminological database of the European Union.
2009: "The Ethnologue" launches its 16th edition as an encyclopedic reference work.

WEBSITES

Alis Technologies: http://www.alis.com/
Aquarius.net: Directory of Localization Experts: http://www.aquarius.net/
ASCII Table: http://www.asciitable.com/
Asia-Pacific Association for Machine Translation (AAMT): http://www.aamt.info/
Association for Computational Linguistics (ACL): http://www.aclweb.org/
Association for Machine Translation in the Americas (AMTA): http://www.amtaweb.org/
CALL@Hull: http://www.fredriley.org.uk/call/
ELRA (European Language Resources Association): http://www.elra.info/
ELSNET (European Network of Excellence in Human Language Technologies): http://www.elsnet.org/
Encyclopaedia Britannica Online: http://www.britannica.com/
Encyclopaedia Universalis: http://www.universalis-edu.com/
Ethnologue: http://www.ethnologue.com/
Ethnologue: Endangered Languages: http://www.ethnologue.com/nearly_extinct.asp
EUROCALL (European Association for Computer-Assisted Language Learning): http://www.eurocall-languages.org/
European Association for Machine Translation (EAMT): http://www.eamt.org/
European Bureau for Lesser-Used Languages (EBLUL): http://www.eblul.org/
European Commission: Languages of Europe: http://ec.europa.eu/education/languages/languages-of-europe/
European Minority Languages (list of the Institute Sabhal Mòr Ostaig): http://www.smo.uhi.ac.uk/saoghal/mion-chanain/en/
Google Translate: http://translate.google.com/
Grand Dictionnaire Terminologique (GDT): http://www.granddictionnaire.com/
IATE: InterActive Terminology for Europe: http://iate.europa.eu/
ILOTERM (ILO: International Labor Organization): http://www.ilo.org/iloterm/
iLoveLanguages: http://www.ilovelanguages.com/
International Committe on Computational Linguistics (ICCL): http://nlp.shef.ac.uk/iccl/
Internet Dictionary Project (IDP): http://www.june29.com/IDP/
Internet Society (ISOC): http://www.isoc.org/
Laboratoire CLIPS (Communication Langagière et Interaction Personne-Système): http://www-clips.imag.fr/
Laboratoire CLIPS: GETA (Groupe d'Étude pour la Traduction Automatique): http://www-clips.imag.fr/geta/
LINGUIST List (The): http://linguistlist.org/
Localization Industry Standards Association (LISA): http://www.lisa.org/
Logos: Multilingual Translation Portal: http://www.logos.it/
MAITS (Multilingual Application Interface for Telematic Services): http://wwwold.dkuug.dk/maits/
Merriam-Webster Online: http://www.merriam-webster.com/
Natural Language Group (NLG) at USC/ISI: http://www.isi.edu/natural-language/
Nuance: http://www.nuance.com/
OneLook Dictionary Search: http://www.onelook.com/
Oxford English Dictionary (OED): http://www.oed.com/
Oxford Reference Online (ORO): http://www.oxfordreference.com/
PAHOMTS (PAHO: Pan American Health Organization): http://www.paho.org/english/am/gsp/tr/machine_trans.htm
Palo Alto Research Center (PARC): http://www.parc.com/
Palo Alto Research Center (PARC): Natural Language Processing: http://www.parc.com/work/focus-area/NLP/
RALI (Recherche Appliquée en Linguistique Informatique): http://www-rali.iro.umontreal.ca/
Reverso: Free Online Translator: http://www.reverso.net/
SDL: http://www.sdl.com/
SDL: FreeTranslation.com: http://www.freetranslation.com/
SDL Trados: http://www.trados.com/
Softissimo: http://www.softissimo.com/

SYSTRAN: http://www.systranlinks.com/
SYSTRANet: Free Online Translator: http://www.systranet.com/
TEI: Text Encoding Initiative: http://www.tei-c.org/index.xml
TERMITE (Terminology of Telecommunications): http://www.itu.int/terminology/index.html
*tmx Vokabeltrainer: http://www.tmx.de/
Transparent Language: http://www.transparent.com/
TransPerfect: http://www.transperfect.com/
Travlang: http://www.travlang.com/
Travlang's Translating Dictionaries: http://dictionaries.travlang.com/
UNDL (Universal Networking Digital Language) Foundation: http://www.undl.org/
Unicode: http://www.unicode.org/
Yahoo! Babel Fish: http://babelfish.yahoo.com/
YourDictionary.com: http://www.yourdictionary.com/
YourDictionary.com: Endangered Languages: http://www.yourdictionary.com/elr/index.html
W3C: World Wide Web Consortium: http://www.w3.org/
W3C Internationalization Activity: http://www.w3.org/International/
WELL (Web Enhanced Language Learning): http://www.well.ac.uk/
Wordfast: http://www.wordfast.org/
Xerox XRCE (Xerox Research Centre Europe): http://www.xrce.xerox.com/
Xerox XRCE: Cross-Language Technologies: http://www.xrce.xerox.com/competencies/cross-language/

End of Project Gutenberg's The Internet and Languages, by Marie Lebert
*** END OF THE PROJECT GUTENBERG EBOOK THE INTERNET AND LANGUAGES [AROUND THE YEAR 2000] ***

1.E.8. You may charge a reasonable fee for copies of or providing access to or distributing Project Gutenberg™ electronic works provided that:

- You pay a royalty fee of 20% of the gross profits you derive from the use of Project Gutenberg™ works calculated using the method you already use to calculate your applicable taxes. The fee is owed to the owner of the Project Gutenberg™ trademark, but he has agreed to donate royalties under this paragraph to the Project Gutenberg Literary Archive Foundation. Royalty payments must be paid within 60 days following each date on which you prepare (or are legally required to prepare) your periodic tax returns. Royalty payments should be clearly marked as such and sent to the Project Gutenberg Literary Archive Foundation at the address specified in Section 4, “Information about donations to the Project Gutenberg Literary Archive Foundation.”
- You provide a full refund of any money paid by a user who notifies you in writing (or by e-mail) within 30 days of receipt that s/he does not agree to the terms of the full Project Gutenberg™ License. You must require such a user to return or destroy all copies of the works possessed in a physical medium and discontinue all use of and all access to other copies of Project Gutenberg™ works.
- You provide, in accordance with paragraph 1.F.3, a full refund of any money paid for a work or a replacement copy, if a defect in the electronic work is discovered and reported to you within 90 days of receipt of the work.
- You comply with all other terms of this agreement for free distribution of Project Gutenberg™ works.

1.E.9. If you wish to charge a fee or distribute a Project Gutenberg™ electronic work or group of works on different terms than are set forth in this agreement, you must obtain permission in writing from the Project Gutenberg Literary Archive Foundation, the manager of the Project Gutenberg™ trademark. Contact the Foundation as set forth in Section 3 below.

1.F.

1.F.1. Project Gutenberg volunteers and employees expend considerable effort to identify, do copyright research on, transcribe and proofread works not protected by U.S. copyright law in creating the Project Gutenberg™ collection. Despite these efforts, Project Gutenberg™ electronic works, and the medium on which they may be stored, may contain “Defects,” such as, but not limited to, incomplete, inaccurate or corrupt data, transcription errors, a copyright or other intellectual property infringement, a defective or damaged disk or other medium, a computer virus, or computer codes that damage or cannot be read by your equipment.

1.F.2. LIMITED WARRANTY, DISCLAIMER OF DAMAGES - Except for the “Right of Replacement or Refund” described in paragraph 1.F.3, the Project Gutenberg Literary Archive Foundation, the owner of the Project Gutenberg™ trademark, and any other party distributing a Project Gutenberg™ electronic work under this agreement, disclaim all liability to you for damages, costs and expenses, including legal fees. YOU AGREE THAT YOU HAVE NO REMEDIES FOR NEGLIGENCE, STRICT LIABILITY, BREACH OF WARRANTY OR BREACH OF CONTRACT EXCEPT THOSE PROVIDED IN PARAGRAPH 1.F.3. YOU AGREE THAT THE FOUNDATION, THE TRADEMARK OWNER, AND ANY DISTRIBUTOR UNDER THIS AGREEMENT WILL NOT BE LIABLE TO YOU FOR ACTUAL, DIRECT, INDIRECT, CONSEQUENTIAL, PUNITIVE OR INCIDENTAL DAMAGES EVEN IF YOU GIVE NOTICE OF THE POSSIBILITY OF SUCH DAMAGE.

1.F.3. LIMITED RIGHT OF REPLACEMENT OR REFUND - If you discover a defect in this electronic work within 90 days of receiving it, you can receive a refund of the money (if any) you paid for it by sending a written explanation to the person you received the work from. If you received the work on a physical medium, you must return the medium with your written explanation. The person or entity that provided you with the defective work may elect to provide a replacement copy in lieu of a refund. If you received the work electronically, the person or entity providing it to you may choose to give you a second opportunity to receive the work electronically in lieu of a refund. If the second copy is also defective, you may demand a refund in writing without further opportunities to fix the problem.

1.F.4. Except for the limited right of replacement or refund set forth in paragraph 1.F.3, this work is provided to you ‘AS-IS’, WITH NO OTHER WARRANTIES OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PURPOSE.

1.F.5. Some states do not allow disclaimers of certain implied warranties or the exclusion or limitation of certain types of damages. If any disclaimer or limitation set forth in this agreement violates the law of the state applicable to this agreement, the agreement shall be interpreted to make the maximum disclaimer or limitation permitted by the applicable state law. The invalidity or unenforceability of any provision of this agreement shall not void the remaining provisions.

1.F.6. INDEMNITY - You agree to indemnify and hold the Foundation, the trademark owner, any agent or employee of the Foundation, anyone providing copies of Project Gutenberg™ electronic works in accordance with this agreement, and any volunteers associated with the production, promotion and distribution of Project Gutenberg™ electronic works, harmless from all liability, costs and expenses, including legal fees, that arise directly or indirectly from any of the following which you do or cause to occur: (a) distribution of this or any Project Gutenberg™ work, (b) alteration, modification, or additions or deletions to any Project Gutenberg™ work, and (c) any Defect you cause.

Section 2. Information about the Mission of Project Gutenberg™

Project Gutenberg™ is synonymous with the free distribution of electronic works in formats readable by the widest variety of computers including obsolete, old, middle-aged and new computers. It exists because of the efforts of hundreds of volunteers and donations from people in all walks of life.

Volunteers and financial support to provide volunteers with the assistance they need are critical to reaching Project Gutenberg™'s goals and ensuring that the Project Gutenberg™ collection will remain freely available for generations to come. In 2001, the Project Gutenberg Literary Archive Foundation was created to provide a secure and permanent future for Project Gutenberg™ and future generations. To learn more about the Project Gutenberg Literary Archive Foundation and how your efforts and donations can help, see Sections 3 and 4 and the Foundation information page at www.gutenberg.org.

Section 3. Information about the Project Gutenberg Literary Archive Foundation

The Project Gutenberg Literary Archive Foundation is a non-profit 501(c)(3) educational corporation organized under the laws of the state of Mississippi and granted tax exempt status by the Internal Revenue Service. The Foundation's EIN or federal tax identification number is 64-6221541. Contributions to the Project Gutenberg Literary Archive Foundation are tax deductible to the full extent permitted by U.S. federal laws and your state's laws.

The Foundation's business office is located at 809 North 1500 West, Salt Lake City, UT 84116, (801) 596-1887. Email contact links and up to date contact information can be found at the Foundation's website and official page at www.gutenberg.org/contact

Section 4. Information about Donations to the Project Gutenberg Literary Archive Foundation

Project Gutenberg™ depends upon and cannot survive without widespread public support and donations to carry out its mission of increasing the number of public domain and licensed works that can be freely distributed in machine-readable form accessible by the widest array of equipment including outdated equipment. Many small donations ($1 to $5,000) are particularly important to maintaining tax exempt status with the IRS.

The Foundation is committed to complying with the laws regulating charities and charitable donations in all 50 states of the United States. Compliance requirements are not uniform and it takes a considerable effort, much paperwork and many fees to meet and keep up with these requirements. We do not solicit donations in locations where we have not received written confirmation of compliance. To SEND DONATIONS or determine the status of compliance for any particular state visit www.gutenberg.org/donate.

While we cannot and do not solicit contributions from states where we have not met the solicitation requirements, we know of no prohibition against accepting unsolicited donations from donors in such states who approach us with offers to donate.

International donations are gratefully accepted, but we cannot make any statements concerning tax treatment of donations received from outside the United States. U.S. laws alone swamp our small staff.

Please check the Project Gutenberg web pages for current donation methods and addresses. Donations are accepted in a number of other ways including checks, online payments and credit card donations. To donate, please visit: www.gutenberg.org/donate

Section 5. General Information About Project Gutenberg™ electronic works

Professor Michael S. Hart was the originator of the Project Gutenberg™ concept of a library of electronic works that could be freely shared with anyone. For forty years, he produced and distributed Project Gutenberg™ eBooks with only a loose network of volunteer support.

Project Gutenberg™ eBooks are often created from several printed editions, all of which are confirmed as not protected by copyright in the U.S. unless a copyright notice is included. Thus, we do not necessarily keep eBooks in compliance with any particular paper edition.

Most people start at our website which has the main PG search facility: www.gutenberg.org.

This website includes information about Project Gutenberg™, including how to make donations to the Project Gutenberg Literary Archive Foundation, how to help produce our new eBooks, and how to subscribe to our email newsletter to hear about new eBooks.