Top languages in global information production

Sergey Lobachev
Casual Reference Librarian
London Public Library, ON

Introduction

The amount of information produced around the world has grown rapidly during recent decades. The advancement of digital technology contributed to this growth by providing a solution for effective storage of large amounts of data. It was estimated that at the beginning of the new millennium information available in print, film, optical and magnetic formats was equivalent to about "250 megabytes for each man, woman, and child on earth" (Varian and Lyman). This statistic may create a false impression that availability of information is also growing. However, there are many barriers which prevent access to information resources. The so-called "Digital Divide" is the most well-known and most discussed issue in library literature, but it is certainly not the only one (Aqili and Moghaddam; James; Berube, etc.).

This paper attempts to examine global information production from a cultural perspective. Its goal is to answer the question: "Which languages are most widely used in the production and dissemination of information?" In other words, if we were to gather all books, journals, films and web pages published and created on the planet, what part of this huge collection would be available in English, French, Spanish, Chinese and other languages? One might agree that English would be at the top of the list, but what language would follow it? What would the top ten languages be? What percentage of overall information resources might each language comprise?

Answering these questions will enable us to better understand the diversity of the information universe and to determine current trends in global information production.

Methods and Data

Information exists in numerous formats. The scope of this research was limited to those information sources which are commonly available through the public domain, i. e. libraries and the Internet. These include books, academic journals, newspapers and popular magazines, films, and web pages. Government documents, archival materials, technical documentation, and computer files were excluded from this analysis, although they constitute the major part of global information resources (Varian and Lyman). The overall goal was not to provide precise and comprehensive data (an enormously difficult task), but rather to give a sense of the kind of information environment in which we are living. Obtaining relative results was more important than calculating statistics.

The first step was to determine the distribution of different languages for each type of information resource. This data was summarized and compared to the percentage of the world's literate population using each corresponding language. Literacy was considered a minimum requirement for accessing and using information, which in most cases is composed of textual characters. Exceptions might include audio, video, and graphic materials. This comparison permitted an estimate of the discrepancy between the population which is potentially capable of using information and the amount of information available in different languages.

The following data were collected for the various information formats.

Books

The UNESCO Institute for Statistics (UIS) remains the only organization which provides relatively reliable numbers about global book publishing (UNESCO Institute for Statistics). The UIS collects publishing data from questionnaires distributed every second year to all member states of UNESCO. Information on non-member states and territories is collected from other sources. The disadvantages of this method are similar to those affecting all survey research. They include non-response, delays, misinterpretation of questions and unavailability of data due to political or economic circumstances (Altbach and Hoshino 165). In general, the response rate tends to be higher for countries with more planned or controlled economies.

The most recent statistics on global book publishing were released in the last edition of the UNESCO Statistical Yearbook (1999). The data was collected by country, using general categories of the Universal Decimal Classification System (UDC). A book was defined as a non-periodical printed publication of at least 49 pages (UNESCO statistical yearbook). The accuracy and completeness of the reported data varies from year to year, due to inconsistent response rates. The most complete data on the number of book titles published worldwide is available for 1995 (total 918,964 titles) (Altbach and Hoshino 164). The language of publication was determined based on the official or widely used language in a particular country (Gordon and Grimes). Table 1 shows the results obtained for book production by number of titles for the top languages around the globe.

Table 1. Book publishing by language.

Language	Number of titles	Percentage of total
English	200,698	**21,84 %**
Chinese (Mandarin)	100,951	**10,99 %**
German	89,986	**9,78 %**
Spanish	81,649	**8,88 %**
Japanese	56,221	**6,12 %**
Russian	48,619	**5,29 %**
French	44,224	**4,81 %**

Korean	35,864	**3,90 %**
Italian	34,768	**3,78 %**
Dutch	34,067	**3,71 %**
Portuguese	33,430	**3,64 %**

Newspapers and magazines

The data about periodicals for a general audience, such as newspapers and magazines, was retrieved from Ulrich's Periodicals Directory. This is the world's largest database of bibliographic and publisher information about more than 300,000 serials of all types. This database allows a search of serial titles by language. If a serial includes an abstract, summaries or section in additional languages, this information is also provided (Ulrich's Periodicals Directory 1: xi). In other words, more than one language could be assigned to one periodical. Many international publications, for example, provide abstracts and summaries in English. Newspapers in minority languages often include articles written in majority languages. Taking into account all periodicals regardless of how often they are published may cause a significant statistical error when retrieving large amounts of data. To reduce potential error, the search strings were limited to daily serials, which in most cases are published in one language.

The type of document was limited to the following categories: "consumer", "trade", "newspaper" and "newsletter/bulletin". The following table represents the data for serials with "active" status in 2007.[1]

Table 2. Newspaper and magazine production by language.

Language	Number of titles	Percentage of total
English	2499	**62.55 %**
Spanish	277	**6.93 %**
German	235	**5.88 %**
Chinese (Mandarin)	156	**3.90 %**
Hindi	117	**2.93 %**
French	95	**2.38 %**
Polish	44	**1.10 %**
Russian	38	**0.95 %**
Italian	36	**0.90 %**
Portuguese	35	**0.88 %**

Scholarly journals

The same method was used for scholarly journals. The document type selected in Ulrich's Periodicals Directory was "academic/scholarly". In this case, all types of serials were considered regardless of frequency of publication. Table 3 shows the number and percentage of scholarly journal titles with "active" status in 2007.[1]

Table 3. Scholarly journal production by language.

Language	Number of titles	Percentage of total
English	28,131	**45,24%**
German	6,848	**11,01%**
Chinese (Mandarin)	4,047	**6,51%**
Spanish	3,522	**5,66%**
French	3,074	**4,94%**
Japanese	2,149	**3,46%**
Italian	1,860	**2,99%**
Polish	1,060	**1,70%**
Portuguese	1,055	**1,70%**
Dutch	922	**1,48%**
Russian	808	**1,30%**

Films and video

The most comprehensive resource on film and video production is the Internet Movie Database (IMDb). It covers more than 1.2 million movies, TV episodes and series. The data is collected through voluntary submissions of information by the people in the film industry and web site visitors. The accuracy of entries is verified by professional staff, who rely on press kits, official biographies, interviews, and on-screen credits (Internet Movie Database).

The sampling was not limited to any specific genre. I assumed that any kind of video production might be considered as a source of information, regardless of intended audience. The database was searched by language of dialogue. Table 4 summarizes the number of film/video titles for the most widely used languages in the industry for the period from 1990 to 2007.[2]

Table 4. Film and video production by language.

Language	Number of titles	Percentage of total
English	158,611	**34,89%**
Spanish	23,256	**5,12%**
German	16,523	**3,63%**
French	15,171	**3,34%**
Japanese	7,811	**1,72%**
Italian	4,927	**1,08%**
Danish	3,967	**0,87%**
Dutch	3,445	**0,76%**
Portuguese	3,213	**0,71%**
Russian	2,715	**0,60%**
Hindi	2,357	**0,52%**

Web pages

Few attempts have been made to estimate language disparities on the Internet. Several reports on Internet statistics by language were released between 1997 and 2004 by Global-reach, an international marketing company; Alis Technologies, a Canadian company; Vilaweb, a Catalan electronic newspaper; and the educational web site Netz-tipp (Murray 35-36; Gorski and Clark 30-34; Web Languages Hit Parade).[3] These reports suggest that the proportion of English content has decreased overtime. In 1997, English web pages comprised 82.3 % of the World Wide Web, but in 2002 they comprised only 56.4 %.

Table 5 represents the latest statistics on language distribution on the Internet. The data was obtained by performing searches in Google and Alltheweb with switched linguistic filters for each respective language (Ebbertz).

Table 5. Distribution of languages on the Internet.

Language	Web pages (millions)	Percentage of total
English	1142,5	**56,43%**
German	156,2	**7,71%**
French	113,1	**5,59%**
Japanese	98,3	**4,86%**
Spanish	59,9	**2,96%**
Chinese (Mandarin)	48,2	**2,38%**
Italian	41,1	**2,03%**
Dutch	38,8	**1,92%**
Russian	33,7	**1,66%**
Korean	30,8	**1,52%**
Portuguese	29,4	**1,45%**

Literate population

To take advantage of the majority of information resources, it is important to understand textual images, in other words, to be literate. Traditionally, "literacy" is defined as the ability to read and to write short simple statements in any language.[4] From this perspective, it would be more appropriate to compare the summarized results of information production with the literate population rather than with the total number of native speakers.

The literacy data was obtained from the CIA statistics and from the "Ethnologue", a catalogue of world languages published by SIL International.[5] Literate population was estimated based on the total population of each country, where a given language is spoken, and the literacy rate in this country (The World factbook; Gordon and Grimes). If more than one language is spoken in the country, the same

literacy rate was applied for each language. When two sources contradicted each other, preference was given to data provided by the CIA.

Table 6. Literate population of the world.

Language	Literate population	Percentage of the world's literate population
Chinese (Mandarin)	794,947,565	**14,68%**
English	572,977,034	**10,58%**
Spanish	295,968,824	**5,47%**
Hindi/Urdu	230,560,488	**4,26%**
Arabic	229,444,922	**4,24%**
French	220,326,329	**4,07%**
Russian	194,503,049	**3,59%**
Portuguese	191,739,619	**3,54%**
Japanese	126,159,159	**2,33%**
Bengali	107,897,009	**1,99%**
German	93,969,555	**1,74%**

Findings and conclusions

The results for information production in different languages are shown in table 7. For each language, the average of the percentage of total information production in every format was calculated and compared to the percentage of the literate population in each corresponding language. As described above, the calculation was based on data derived from different sources, the accuracy and quality of which may vary depending on methods used for data collection. There are also chronological gaps. Not all data is available for the same periods of time.

Does it make the calculation unreliable? The answer depends on the purpose for which the statistics are to be used. I was primarily interested in finding comparative data rather than accurate numbers. My goal was to determine the most common languages in global information production and the proportion of information resources available in those languages. The possible errors do not significantly change the final conclusion.

Table 7. Information users and information production in most spoken languages.

Language	Literate population	Information production
English	10,58%	**44,29%**
German	1,74%	**7,60%**
Spanish	5,47%	**5,91%**
Chinese (Mandarin)	14,68%	**4,85%**
French	4,07%	**4,21%**
Japanese	2,33%	**3,34%**
Italian	1,09 %	**2.16%**
Russian	3,59%	**1,96%**
Portuguese	3,54%	**1,68%**
Dutch	0,43 %	**1.67%**
Korean	1,36%	**1,20%**
Hindi	4,26%	**0,96%**
Arabic	4,24%	**0,43%**
Bengali	1,99%	**0,12%**

As can be seen from Table 7, almost 78 % of all information in the world is produced in the following ten languages: English, German, Spanish, Chinese (Mandarin), French, Japanese, Italian, Russian, Portuguese, and Dutch. English dominates universal information space and constitutes more than 44 % of printed and electronic materials. German follows English and comprises 7.6 % of the global information production.

14.69 % of the world population is literate in Mandarin, the most spoken language in the world, but only 4.85 % of global information resources are produced in this language.

Other widely spoken languages include Hindi, Arabic, Bengali, and Korean. At the same time, the number of information resources in these languages is relatively small. For example, there are 230 million people literate in Arabic, which constitutes 4.24 % of the world's entire literate population, but only 0.43 % of all information is available in the Arabic language.

What do these numbers mean?

First of all, they can measure the importance of a particular language. Its rank is not necessarily related to a percentage of the literate population, but rather depends upon the level of cultural and economic development of the countries where the language is used.

Secondly, they underline the gap between the users of information and available information resources. They clearly show how the "language divide" contributes to the exclusion of countries and peoples from universal knowledge. This primarily

concerns countries with low literacy rates and poor education. At the same time, the educated community tends to view English as a universal language. Many countries have special programs which encourage citizens to achieve proficiency in English (Weber).[6]

Nevertheless, we need to realize that more than half of the world's information resources are produced in non-English languages. These resources will likely continue to grow in the near future. The "Global Trends 2025" report, recently released by the US National Intelligence Council, projected the increasing role of Brazil, Russia, China and India in the world economy (BBC News). If this forecast is true, we may expect the rise of information production in Portuguese, Russian and Chinese.

This trend must be taken seriously by publishers and vendors in English-speaking countries, where non-English resources are largely ignored. According to the Bowker publishing group, only 3 % of all books available for sale in the United States are new translations from other languages (English-Speaking Countries). The term "language divide" can be equally applied to the English-speaking world.

There are many opportunities for librarians to respond to the challenges of multilingualism in information production. One of them is by providing equal access to information resources regardless of the language of origin. Today, however, most widely-distributed indexing databases cover primarily English content. Non-English materials are not fully searchable, and access to full-text electronic articles from non-English periodicals is not always available.

Future work

This paper has outlined the language profile of global information production. It should be considered a first step toward further research, which may focus on the following aspects:

- Finding more comprehensive and more reliable statistics for information production in the most widely used languages.
- Widening the scope of the research by taking into account as many formats of information as possible (archival documents, government publications, digital resources, audio recordings, etc.).
- Broadening the definition of literacy and examining the role of multilingualism in literacy and information production.
- Finding historical data, which helps us understand the evolution of information production in different languages during the last decades, or centuries. These statistics also help to determine if there is a trend towards a decline in the prevalence of the English language in various formats of information resources.
- Qualitative analysis of the content of information produced in different languages.
- Comparative analysis of languages of library collections around the world.

Acknowledgements

The early version of this research was presented as a poster session at the CLA Conference in Vancouver in May 2008. I would like to thank Edwin Perry, Liaison Librarian at the University of Regina, and two anonymous reviewers for their comments and suggestions.

Works Cited

Altbach, Philip G. and Edith S. Hoshino. International Book Publishing: An Encyclopedia. Garland reference library of the humanities, vol. 1562. New York: Garland Pub, 1995.

Aqili, Seyed V., and Alireza I. Moghaddam. "Bridging the digital divide: The role of librarians and information professionals in the third millennium". Electronic Library 26.2 (2008): 226-37.

BBC News, Americas. "US Global Trends report: Key points". 21 November 2008. <http://news.bbc.co.uk/2/hi/americas/7741241.stm>.

Berube, Linda. "The Digital Divide, or Who Gets to Be Part of the Information Society". Multimedia Information and Technology. 32.3 (2006): 86-9.

Ebbertz, Martin. Das Internet spricht Englisch ... und neuerdings auch Deutsch: Sprachen und ihre Verbreitung im World-Wide-Web, 2002. 30 November 2008. <http://www.netz-tipp.de/sprachen.html>.

"English-Speaking Countries Published 375,000 New Books Worldwide in 2004. News Release", Bowker: New Providence, N. J., 2005. 20 April 2008. <http://www.bowker.com/press/bowker/2005_1012_bowker.htm>.

Gordon, Raymond G., and Barbara F. Grimes. Ethnologue: Languages of the World. Dallas, Tex: SIL International, 2005. <http://www.ethnologue.com>.

Gorski, Paul and Christine Clark. "Multicultural Education and the Digital Divide: Focus on Language". Multicultural Perspectives, 4.2 (2002): 30-34.

The Internet Movie Database. 18 April 2008. <www.imdb.com>.

James, Jeffrey. "Digital Preparedness Versus the Digital Divide: A Confusion of Means and Ends". Journal of the American Society for Information Science and Technology. 59.5 (2008): 785-91.

Murray, Denise E. "New Frontiers in Technology and Teaching", in Davison, Chris., ed. Information Technology and Innovation in Language Education, Hong Kong: Hong Kong University Press, 2005.

OCLC. Web Characterization. 20 April 2008. <http://www.oclc.org/research/projects/archive/wcp>.

O'Neill, Edward T., Lavoie, Briann F. and Rick Bennett. "Trends in the Evolution of the Public Web: 1998--2002". D-Lib Magazine, 9.4 (2003). <http://www.dlib.org/dlib/april03/lavoie/04lavoie.html>

Ulrich's Periodicals Directory, 2007: Including Irregular Serials & Annuals. New Providence, N.J.: CSA, 2006. vol. 1-2. 18 April 2008. <http://www.ulrichsweb.com/ulrichsweb>.

UNESCO Institute for Statistics. <http://stats.uis.unesco.org>.

UNESCO. UNESCO statistical yearbook 1999. Paris: UNESCO, 1999. <http://www.uis.unesco.org/statsen/statistics/yearbook/YBIndexNew.htm>.

Varian, Hal R., and Peter Lyman. How Much Information? Berleley, Calif: School of Information Management and Systems at the University of California at Berkeley, 2000. Executive summary. 18 April 2008.

<http://www2.sims.berkeley.edu/research/projects/how-much-info/summary.html>.

"Web Languages Hit Parade", Alis Technologies, Inc., 1997. 30 November 2008. <http://alis.isoc.org/palmares.en.html>.

Weber, George. "Top languages: The World's 10 most influential languages". Language Today. 3 (1997): 12-18. 20 April 2008. <http://www.andaman.org/BOOK/reprints/weber/rep-weber.htm>.

The world factbook. Washington, D.C.: Central Intelligence Agency: Supt. of Docs., 1981. 20 April 2008. <https://www.cia.gov/library/publications/the-world-factbook>.

Notes

[1] The data was retrieved in April 2008.

[2] The data was retrieved on 18 April 2008.

[3] The Vilaweb data is presented on the web site of the ClickZ Network, which is specialized on providing electronic marketing news, information, commentary, research, and reference services. See <http://www.clickz.com/showPage.html?page=408521>. The original Vilaweb report is no longer available online, however, it is often cited in scholarly literature. The data collected by Vilaweb is very close to the results obtained by OCLC, which conducted the Web Characterization research project in 1998-2002. The goal of the project was to analyze the size and content of the Web based on samples of publicly available resources. (O'Neill, Lavoie and Bennett; OCLC).

[4] This definition is used by UNESCO for collection literacy data around the world. See <http://www.uis.unesco.org/ev.php?ID=5013_201&ID2=DO_TOPIC>.

[5] The data was retrieved in April 2007. SIL International (Summer Institute of Linguistic) is a faith-based organization that studies, documents, and assist in developing of world's lesser-known languages. Its premier publication covers more than 6,900 living languages (Gordon and Grimes).

[6] George Weber ranked the top ten most influential languages as follows: English, French, Spanish, Russian, Arabic, Chinese, German, Japanese, Portuguese, Hindi. The results were based on the analysis of six main factors: number of primary speakers, number of secondary speakers, economic power of countries using the language, number of major areas of human activity in which the language is important, number of population of countries using the language, and socio-literary prestige (Weber).