



United Nations
Educational, Scientific and
Cultural Organization

Twelve years of measuring linguistic diversity in the Internet: balance and perspectives

Twelve years of measuring linguistic diversity in the Internet: balance and perspectives

By Daniel Pimienta, Daniel Prado and Álvaro Blanco



UNESCO publications for the
World Summit on the Information Society
Communication and Information Sector



United Nations
Educational, Scientific and
Cultural Organization

Twelve years of measuring linguistic diversity in the Internet: balance and perspectives

By

Daniel Pimienta, Daniel Prado and Álvaro Blanco

Disclaimer

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion what so ever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The authors are responsible for the choice and the presentation of the facts contained in this book and for the opinions expressed therein, which are not necessarily those of UNESCO and do not commit the Organization.

Recommended catalogue entry:

Twelve years of measuring linguistic diversity in the Internet: balance and perspectives

Edited by the Information Society Division, Communication and Information Sector, UNESCO, 58 p., 21 cm.

Co-editor: Viesti Dickens

I – **Twelve years of measuring linguistic diversity in the Internet: balance and perspectives**

II – **Daniel Pimienta, Daniel Prado and Álvaro Blanco**

III – Languages, Internet, Web, linguistic diversity, linguistic policies, indicators

Published in 2009

By the United Nations Educational, Scientific and Cultural Organization
7, place de Fontenoy, 75352 Paris 07 SP, Paris, France

© UNESCO 2009

All rights reserved

CI-2009/WS/1

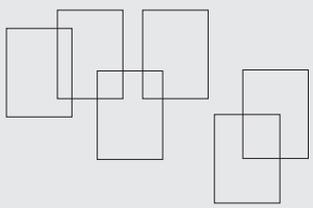
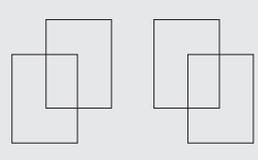
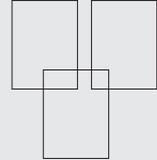
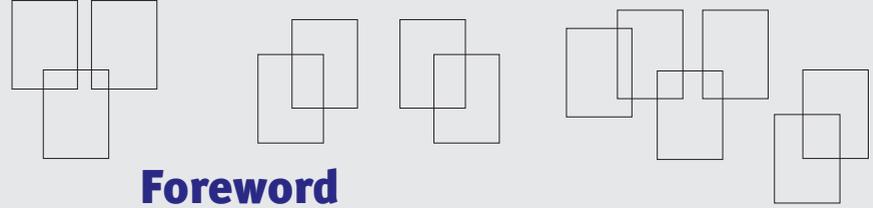


Table of content

FOREWORD	i
ACKNOWLEDGEMENTS	iii
ABSTRACT	v
1. INTRODUCTION	1
1.1 BIRTH OF A PROJECT	1
1.2 OBJECTIVES OF THE PAPER	3
2. CONTEXT OF A PROJECT	5
3. HISTORY OF A PROJECT	9
4. METHODOLOGY	11
4.1 LINGUISTIC METHODOLOGY	13
4.2 SEARCH ENGINE METHODOLOGY	20
4.3 STATISTICAL METHODOLOGY	23
4.4 INDICATORS BUILDING	24
5. RESULTS	29
5.1 MAIN RESULTS	29
5.2 ANALYSIS PER COUNTRY	34
5.3 OTHER SPACES FOR LANGUAGE DIVERSITY	38
5.4 CULTURAL DIVERSITY	39
6. EVALUATION OF THE METHOD	43
6.1 ITS UNIQUENESS AND ADVANTAGES	43
6.2 ITS WEAKNESSES AND LIMITATIONS	44
7. EVALUATION OF OTHER METHODS	45
8. PERSPECTIVES	55
REFERENCES	57

A decorative graphic at the top of the page consists of several overlapping squares of varying sizes and positions, creating a geometric pattern. The squares are white with black outlines. Below this graphic, the word "Foreword" is written in a bold, blue, sans-serif font.

Foreword

The Internet is a major opportunity to improve the free flow of information and ideas throughout the world. Committed to building inclusive Knowledge Societies, UNESCO is actively engaged in efforts to improve cultural and linguistic diversity on the Internet and broaden access to information for all.

Over the last year, the celebration of the International Year of Languages by UNESCO has attracted the attention of policy-makers around the world and larger section of public opinion to the strategic relevance of languages and linguistic policies for development. With the introduction of Internationalized Domain Names (IDNs) in Internet addresses, the issue of Internet access in local scripts and languages has further become front and center in the recent debates on the Internet Governance.

Today, the international community is increasingly interested to enable a greater number of people to access and use the Internet in their own scripts and languages. The relationship between languages on the Internet and diversity of language within a country indicates that countries have an important role to play in adopting an appropriate linguistic policy for the Internet. Such a comprehensive linguistic policy requires a specific component to address linguistic diversity in the virtual world, as well as relevant figures based on reliable indicators quantifying the situation.

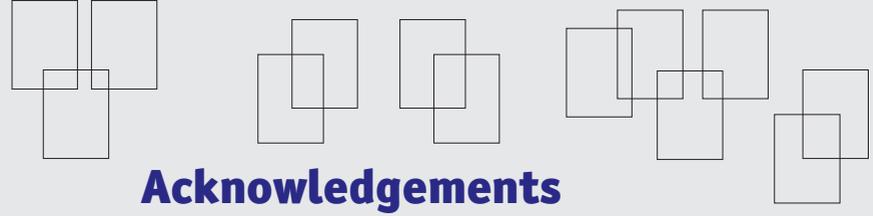
To this end, UNESCO asked the experts from FUNREDES and Union Latina to update the study entitled "Measuring Linguistic Diversity on the Internet" and published for the World Summit on the Information Society in 2005. UNESCO is committed to an approach of statistics and measurements that goes beyond a techno-centric view to consider the importance of content and the enabling environment, while at the same time acknowledging limitations in measuring culture and content represented on the Internet.

This study presents a variety of methods used over 12 years to measure linguistic diversity on the Internet. The results of the study

dispel some of the myths surrounding existing figures, for example, the dominant presence of English on the Web.

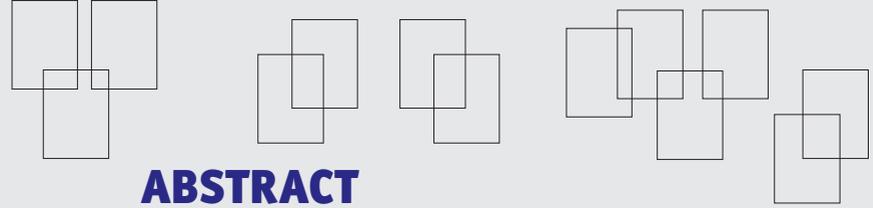
I hope that the study will be widely used, especially by policy makers and professionals assigned to leadership responsibility for introducing, applying and evaluating linguistic policies, strategies, programmes and projects in the Internet. It should also be useful to academics involved in research to measure linguistic diversity in the Internet. I recommend this publication to them all. I hope that this publication, consistent with the UNESCO Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace adopted in 2003, will facilitate the formulation of linguistic policies conducive to cultural and linguistic diversity on the Internet.

Abdul Waheed Khan
Assistant Director-General for
Communication and Information
UNESCO



Acknowledgements

The authors of this publication are Daniel Pimienta (Head of FUNREDES, Member of Executive Committee of MAAYA network, Researcher at Université Antilles-Guyane in Martinique), Daniel Prado (Head of the Department of Terminology and Language Industries of Union Latine and Member of Executive Committee of MAAYA network) and Àlvaro Blanco (Head of FUNREDES, Spain). Several other people from Union Latine and FUNREDES also participated in this project, including Marcelo Sztrum, who managed the team of linguists for the word-concepts selection, Benoit Lamey, who wrote the computer programme that has entailed limited human intervention in the measurement process, and Roger Price, who provided essential inputs to the statistical part of the research.

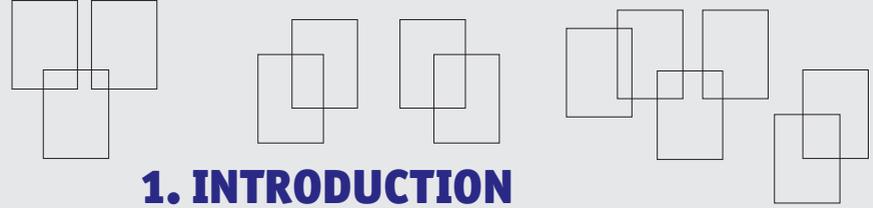
The top of the page features a decorative header with the word "ABSTRACT" in a bold, blue, sans-serif font. Above the text is a series of overlapping, light gray squares of various sizes and orientations, creating a geometric pattern. The squares are arranged in a way that they appear to be floating or layered, with some partially obscuring others. The background of the header is a light gray color.

ABSTRACT

FUNREDES and Union Latine have designed an original research method to measure linguistic diversity in cyberspace. The aim was to use search engines and a sample of word-concepts to measure the proportionate presence of these concepts in their various linguistic equivalences (in Latin languages, English and German) in cyberspace. The research, undertaken from 1996 to 2008, has enabled interesting indicators to be built in order to measure linguistic diversity. Additionally, some basic evaluations of the cultural projections associated with these languages (mentioned above) were undertaken.

This paper describes the research method and its results, advantages and limitations. It also provides an overview of existing alternative methods and results, for comparison. The paper concludes with the examination of different perspectives in a field which have in the past been considered to have been characterized by a lack of scientific rigor. This has led to some misinformation about the dominant presence of English on the Web. It is a topic that is only now slowly attracting due attention from international organizations and the academic world.

All relevant detailed data about the methodology and results of this study are freely available on the Web. Several publications have been made on the basis of the research results that document the project chronologically. This has led to some difficulties for readers to obtain a complete overview of the project in a single document. This paper attempts to solve that issue by providing a complete description of the whole process and results in a single document. It references previous reports, which could be useful for researchers or policy makers interested in deepening their knowledge of the method or results. The paper also aims to contribute to the sensitization of civil society to the theme of linguistic diversity in cyberspace.



1. INTRODUCTION

1.1 BIRTH OF A PROJECT

In December 1995, during the Francophone Summit in Cotonou, the presence of English in the recently born World Wide Web was publicly quoted as being above 90%. This figure provided the impetus for statements criticising the Internet for its inherent linguistic bias, triggering a reaction from FUNREDES in defence of the Internet. The FUNREDES team first tried, unsuccessfully, to locate the source for this figure that connotes English dominance on the Internet, and then looked for a way to produce some initial rough figures. The idea to harness the power of search engines for the research (a world dominated by Altavista at that time) was then born. A first trial was established to obtain an approximate idea of the split of English, French and Spanish on the Web¹. Additionally, another rough estimate of the representation of the cultures associated with those languages was made, by weighting the presence in the Web of names of important personalities of different categories and comparing them².

Both processes, and especially that of linguistic measurement, obviously lacked any kind of scientific value at this stage. However, they provided with the opportunity to:

- 1) make a very rough estimate of the presence of English on the Web as being around 80%;
- 2) gauge the challenges to be overcome in order to obtain a reliable method of measurement of linguistic diversity on the Web, based on search engine counts of the number of Web pages featuring a given word;
- 3) show that the global nature of the Web was allowing a fair representation of French-related cultures (at least when clearly disassociated from commercial realms), although less so for

1 <http://funredes.org/lc2005/english/L1.html>

2 <http://funredes.org/lc2005/english/C1.html>

Spanish-related cultures at this time (the situation has significantly evolved since then).

Additionally, this research into online linguistic diversity may have initiated a process of capturing information of interest for Internet documentation and archiving. It has also contributed to the analysis of the historical behaviours of search engines.

In any case, the results paved the way for what became the only existing series of repeated and coherent measurements of the presence of a subset of languages on the Web and other online spaces³. From 1996 to 2008, a transparent presentation of methodology and results was achieved. During this 12 year research period, the predominance of English continued to be overstated as being 80%, despite the formidable speed of the evolution of Internet demographics, which showed a drop from 80% to 40% in terms of English-speaking Internet users⁴. This disinformation was a major challenge for the research team to overcome.

Undertaking this research was not a question of fighting to defend French, but rather, in keeping with FUNREDES' role as an 'ICT4D-focused NGO'⁵, it was a plea for the production of local content. Throughout the research period, the broad perception was that of a massive, pervasive and stable English dominance on the Internet, in the context of a virtual world that was supposed to reflect the linguistic and cultural diversity of the 'real' world. This perception seemed to conspire against the obvious need for a coherent policy for the creation of Web pages in 'mother tongues' and for local content development.

In 2005, UNESCO published a report (see [Measuring linguistic diversity on the Internet](#) in references) which sought to provide an overview of different academic and research perspectives about linguistic diversity on the Internet. The report also included figures

3 In this document, words or expressions such as online, virtual world, Internet, Net or cyberspace are synonymous; The word Web specifically is used to refer to a subset of the Internet (as well as Newsgroup, Blog or Wikipedia for other subsets).

4 GlobalStat reference below.

5 ICT4D stands for Information and Communication Technologies (ICT) for Development. Non Governmental Organizations (NGOs) working in that field try to use ICT to empower persons, communities and countries to change positively their socio-economic conditions.

relating to the online presence of English. To provide some examples, Paollillo's paper - supporting 80% as an accurate calculation for the English presence on the Internet - comprised the main reference of the work from On Line Computer Library Centre (OCLC)⁶ team. Pimienta coordinated a set of papers from researchers from around the world⁷ and argued for an English presence of around 50%. The publication of this report may have marked a historic turning point, leading to more open views about linguistic diversity online and generating further research interest in this important but neglected field of studies.

1.2 OBJECTIVES OF THE PAPER

As mentioned, this paper is primarily an attempt to provide a complete overview and analysis of the results of a series of studies conducted by FUNREDES and Union Latine from 1996 to 2008. Although the methods and results were published with complete transparency over the course of the project, they were presented as a series of events described sequentially. This forced the reader to follow the chronology in order to understand the work. It fell short of providing a clear, coherent and pedagogic report of the work.

This paper rectifies that shortfall. It is the first attempt to synthesise and analyse the results produced by the series of studies. In this paper the values and limitations of the methodology and results are shown and other research projects are analyzed and their limitations exposed. The paper will also explain why and how the used methodology is no longer viable, due to the recent evolution of search engines. This has driven the researchers to consider the need for more ambitious tools to better reflect the reality of the whole Web. Thus the second objective of the paper is to examine the state of a new discipline which is part of *cybermetrics*.

This will provide policy makers around the world with a vision of the evolution and trends of languages on the Internet. It will also provide comprehensive material for researchers or policy makers interested in the area of linguistic diversity on the Internet, at a time when

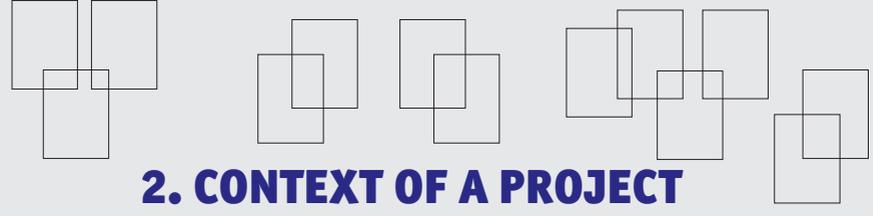
6 <http://www.oclc.org/>

7 Some fine papers which have not been selected in the final report due to a lack of space, can be viewed at: <http://funredes.org/lc/english/unesco/>

the issues are finally receiving the due attention they deserve⁸. In particular, this paper intends to definitively correct the misinformation about the extent of English dominance on the Web. In conclusion, the authors will explain their plans to continue to measure linguistic and cultural diversity on the Internet, for which they are publicly seeking cooperation and support.

Although some parts of the paper may require technical knowledge (linguistic, statistical or that related to the Internet itself) in order to be fully understood, the paper provides an overview that is also relevant for interested non-specialists. The last objective is therefore to stimulate and raise awareness of “netizenship”, so that civil society better understands the importance of linguistic diversity in cyberspace.

8 As witnessed in the Internet Governance Forum in Rio de Janeiro (2007), where a panel coordinated by Brazilian Minister of Culture, Gilberto Gil and with the presence of MAA YA’s President, Adama Samassekou, was dedicated to the theme. See <http://www.intgovforum.org> and especially this link: http://www.intgovforum.org/Rio_Meeting/IGF2-Diversity-13NOV07.txt



2. CONTEXT OF A PROJECT

Linguistic experts generally differ in opinion when giving demographic figures on languages. Definitions and boundaries are complex and reaching a consensus is not easy. The following information will principally draw on David Crystal (see [Language and the Internet](#) in references) as its main source. It should also be noted that for matters specifically related to Latin languages, Union Latine will be the reference for this paper.

The number of languages which have been created by human beings is estimated to be around 40,000. Among them, it is estimated that only between 6,000 and 9,000 are still in use (figures varying depending on the source), and some sources stipulate that an average of one language is lost every two months.

In this context, preserving linguistic diversity becomes an important issue to address. Yet the question naturally arises as to whether the Internet heralds an opportunity or a threat for linguistic diversity.

The answer is not a simple matter and it appears to really depend on several parameters of the language in question: is it local, national, international or a *lingua franca*? Is it from a developed country? Is there a linguistic policy? Is there a linguistic policy thought for cyberspace?

To summarize the situation, a simple language classification table is provided below. It does not pretend to offer any definitive answers. Rather, its purpose is to initiate reflections on a subject that civil society has not been as exposed to as it has for biological diversity, although there are obvious correlations between both matters⁹. Given the current situation of the planet, the lack of policy for protection against a reduction in biological diversity could harm the collective future. The same question could be asked for cultural diversity, and warrants the team attention. The table explicates the implications of exposure to and presence on the Internet for each language category.

⁹ The correlation of the regions in the world with high/low biological diversity with the number of spoken languages shows questioning evidences.

Table 1: Categorization of languages for cyberspace policies requirements

CATEGORY	DOES THE INTERNET HERALD AN OPPORTUNITY?
Main spoken languages ¹	The Internet could increase the online presence of these languages, especially during a transition period when the repartition of Internet users by language is not even due to the digital divide. Note: the thesis here is that this transitory period has been over for the English language as of a few years ago.
Official languages covering more than one developed country (like Italian or Dutch)	There is an opportunity to be seized in the virtual world. The “international” status of these languages shall facilitate trust between speakers to create easy cross-border relations.
Official languages spoken in only one developed country (like Norwegian, Greek, Danish or Japanese)	There is a need for a vigorous virtual linguistic policy to support a presence in the virtual world comparable or stronger than that in the real world. Despite having a sense of longevity in relation to the place of such a language in the world, its speakers may feel a barrier for international relations.
Local languages of developed countries (like Sardinian, Galician, Welsh, or Frisian)	These languages are threatened by pressure from both English and their respective national languages. The diagnostic is uncertain without a virtual linguistic policy. Each case varies and depends on specificities, although the case of Catalan is to be followed as a success story, both at virtual and non virtual level.
Lingua franca of speakers of some developing countries (like Hausa, Quechua, Pulaar or Swahili)	A positive future shall be possible where the digital divide is really overcome and virtual linguistic policies are defined.
Languages of a developing country, that actually cover more than one country, but are only used by native speakers (like Aymara, Guarani or creoles)	Theoretically, a positive future should be possible where the digital divide is really overcome. However, there is a presently a correlation between lack of access to computers and the issue of belonging to indigenous communities, which does not give any sign of changing any time soon. The case of Paraguay where Guarani is given resources following its declared status as an official language is to be followed with interest.

CATEGORY	DOES THE INTERNET HERALD AN OPPORTUNITY?
Official languages of a unique developing country (like Slovenian or Albanese)	They are under strong pressure from both English and respective powerful regional languages, which could trigger negative prospects in the absence of a virtual policy.
Local languages of developing countries (like Chabacano, Maya or Mapuche)	If the language is provided with the appropriate linguistic tools (and first a normalized and stable system for writing and grammar), a linguistic policy focusing the production of local content could help. However there are not many examples today of this kind.
Languages at risk of disappearing (like Ainu.)	The Internet could, at worst, become a formidable tool for conservation of the written or oral patrimony; or at best, accelerator of policies for language adaptation.
Languages very seriously at risk of disappearing (like Yagan)	The Internet could at least allow preservation of the patrimony of that language, if digitalization is undertaken soon enough.

The main message arising from this table is the need for language policies to be established, both in the real world (and Catalan is a good reference for a success story to be studied in this sense) and in the virtual world (where analysis of the actions of *Organisation de la Francophonie* is of interest, as the studies show the positive results obtained from a voluntary policy for content production in French).

In order to create a meaningful linguistic policy, the first step is to obtain relevant figures quantifying the situation, so as to be able to assess and follow up the effects of the policy, based on reliable indicators. Nowadays, a comprehensive policy for language must include a specific component for the virtual world, implying different dynamics, logic and rules in comparison to the real world. The need for reliable data on the presence of languages in the Internet naturally follows.

In that regard, the situation has been paradoxical. The history of the Internet is closely linked to the history of research and the academic world. However, with the birth of the Web and the growth of the commercial part of the Internet, the academic sector has partly given up the creation of Internet demographic data to the private sector, and perhaps more controversially to the marketing sector. This has created privately held, rather than publicly available, data. This has

often led to the lack of transparency of research methodologies. Published figures have not always been produced using scientific criteria. Furthermore, such data production may have been driven by particular commercial or other interests that can influence results obtained or seek to elicit particular information.

In an area for which demographics have been changing at a speed without precedent in human history, this has enabled the creation of myths, like the overwhelmingly dominant and stable presence of English on the Web as being around 80%. This myth has largely lacked a critical response from the academic world.

However, this era seems to have come to an end, as shown by the following facts:

- 1) a coherent series of actions organized and followed by UNESCO¹⁰;
- 2) the emergence of MAAAYA¹¹ from the World Summit of Information Society process; and
- 3) the Language Observatory Project¹², launched by a network of universities.

There is therefore a growing interest from policy makers and academics to take back control and contribute meaningfully to the emerging and necessary linguistic policies for the virtual world, based on the use of reliable indicators.

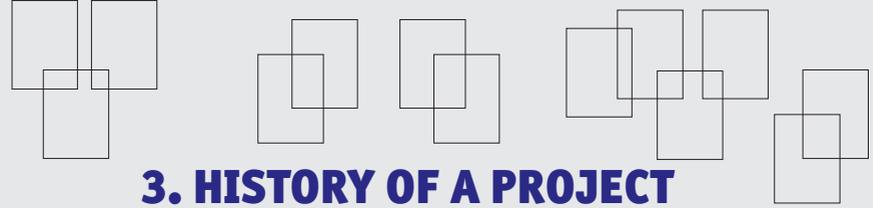
In that historical context, this project can be seen as a unique, pioneering and committed research-action attempt by civil society to resist the influence of disinformation about the Internet. Indeed, linguistic diversity on the Internet is strategic because it directly relates to the issues of digital or knowledge divide, and Internet governance¹³.

10 http://portal.unesco.org/education/en/ev.php-URL_ID=19741&URL_DO=DO_TOPIC&URL_SECTION=201.html

11 World Network for Linguistic Diversity: <http://maaya.org>

12 Led by Professor Mikami, at Nagaoka University of Technology, the Language Observatory Project (LOP) constitutes a world-wide consortium of partners. See <http://www.language-observatory.org/>

13 The Internet Governance Forum is now focusing with intensity the question of Internationalized Domain Names while this is just the tip of the iceberg of linguistic diversity in the Net. See Comment assurer la présence d'une langue dans le cyberspace? in references.



3. HISTORY OF A PROJECT

The history of this research project has been documented on two websites. For those who are interested in following the evolution of this project from 1996 to 2005, the project's initial, historical Web page¹⁴ provides all necessary information, listed as a chronological series of measurements. The second site¹⁵ presents the results after 2005.

The following table summarises the project process across these periods.

Table 2: The series of measurements and steps of the project

DATES	LINGUISTIC STUDY	ENGLISH IN THE WEB / Search Engine	CULTURAL STUDY
6/96	L1: Very rough linguistic results - English, French, and Spanish. - the Web	~80% Altavista	C1: First cultural result
3/97	L2: Repetition of L1	~80% Altavista	
3/98	L3: Repetition with a larger sample - Method of the complement of the empty space - Analysis of Alis Method. - Decision to invest in a solid method in partnership with Union Latine	~80% Altavista	
9/98	L4: first study made with reliable methodology, in partnership with Union Latine and with the financial support of Agence de la Francophonie. - Addition of Italian, Portuguese and Romanian - Addition of Usenet - Start creating linguistic indicators.	75% Hotbot Dejanews	C2: Second cultural results with several improvements of the sample and of the classification scheme. Notable progress of the presence of French and Spanish personalities.

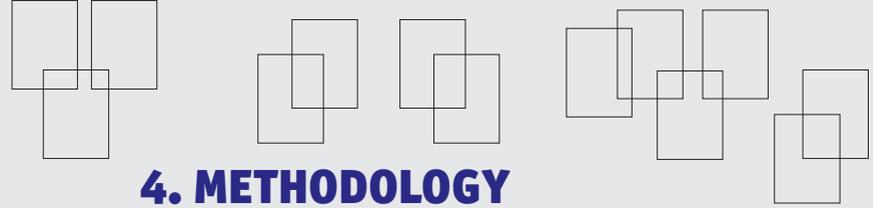
14 <http://funredes.org/lc2005/>

15 <http://funredes.org/lc/>

DATES	LINGUISTIC STUDY	ENGLISH IN THE WEB / Search Engine	CULTURAL STUDY
8/00	L5: Second study made with reliable methodology, in partnership with Union Latine. Creation of a program to automatically generate the whole process from search engine requests and counting to statistical results. Addition of German	60% Google Fast	
1/01 6/01 8/01 10/01 2/02 2/03 2/04 5/04 3/05	L6 ² : - Series of measurements without change of method. - New indicators per country and language for French. (2002) - New indicators per country and language for Portuguese. (2003) - New indicators per country and language for English. (2004)	From 55% To 47% Fast Yahoo Google	C3: September 2001
10/05	- Measurements without change	45% Google	New result on culture
3/06	- Measurements without change	45% Google	
12/07	- Addition of Catalan	45% Yahoo	
5/08	- Measurements without change	Yahoo	New result on culture

September 1998 represents the start of the use of reliable methods and results of the study. With the introduction of a PHP¹⁶ based program for the automation of the whole process, including maintenance of a database of results, September 2000 marks the start of a professional and systematic management of the project.

¹⁶ PHP is a scripting language suited for Web development.



4. METHODOLOGY

The defined methodology is based on a combination of:

- the use of the number of occurrences of each word-concept per language as measured by search engines¹⁷,
- a sample of word-concepts in a given selection of languages,
- a set of standard statistical tools.

The search engines were selected on the basis of meeting a minimum set of specific criteria designed for the study, such as:

- provides reliable figures for counting,
- enables fair treatment of diacritics¹⁸,
- covers the largest possible part of the analyzed space of the Internet.

The word-concept samples to be used for and counted by selected search engines were chosen for their conceptual congruence among the languages of the study, in terms of:

- a perfect syntactic equivalence,
- the best possible semantic equivalence,
- the least possible cultural bias.

The compilation of the Web pages count for each word-concept (the computed sum of the results of the different words associated with each concept¹⁹) is treated as a random variable whose distribution is

17 The search engines output the number of Web pages containing a given word or expression.

18 Diacritics are present in most of the language using Latin alphabet, but not so in English. They allow often to identify different meanings (caña in Spanish has a different meaning as of cana, côte in French is different of cote or of côté). The Internet used at the beginning not to allow codification of diacritics as a consequence of using a character-encoding scheme based on the English alphabet (ASCII - American Standard Code for Information Interchange), with only 7 bits thus permitting no more than 128 different characters.

19 And, in certain cases, providing corresponding corrections.

statistically processed (by average, variance and confidence interval, using the Fisher Law).

The objective is to produce an estimation of the relative weighting of the language in question compared to English, as it is measured in the index²⁰ of the selected search engine. Under certain circumstances (the size of the index being the key factor), it is reasonable to extrapolate the result as a fair representation of the division of languages on the (visible) Web²¹.

In order to obtain an absolute percentage value for the studied languages, as measured in the selected space of the Internet, the absolute weighting of English must first be determined, to serve as a point of reference and comparison. Unfortunately, the research method used did not enable this. Instead, this weighting had to be determined using an additional manual step, undertaken by integrating information drawn from different sources, together with an estimation of the relative weighting of the rest of the languages not included in the study. Periodically repeating the process allowed the researchers to obtain a vision of the evolution of the presence of languages over time.

Although the Web was the main object of the study, other parts of cyberspace were also studied, such as newsgroups, or more recently, various blogs and Wikipedia.

The research method also included consideration of the following:

- the precise criteria for validation and use of search engines;
- the linguistic criteria used to build sample vocabulary (and the corrections which were required in certain situations);
- the statistical tools which were applied to reach final results;
- the building of indicators from these results;
- the building of more complex indicators from more sophisticated use of the method;
- the nature, significance and limitations of the obtained results.

20 By "index of the Search Engine" it is meant here the whole set of the Web pages indexed by the search engine.

21 The invisible Web (also called deep Web) is the sum of dynamic pages produced by data bases or other programmed mechanisms for dynamic pages. Some authors estimated it could be 100 and 500 times larger than the visible Web (see [White Paper: The Deep Web](#) in references).

4.1 LINGUISTIC METHODOLOGY

The list of the 57 concepts (in English) which were used for the language comparison is contained in the following box:

Table 3: List of word-concepts

ambiguity, causality, cheese, compatibility, contiguity, dangerous, December²², density, disparity, divisibility, elasticity, electricity, February, femininity, fertility, fidelity, fraternity, Friday, heterosexuality, homosexuality, horse, humidity, illness, immortality, immunity, incompatibility, infallibility, inferiority, infidelity, instability, inviolability, irregularity, irresponsibility, June, knee, knife, lung, masculinity, Monday, October, parity, equality, probability, productivity, puberty, responsibility, sexuality, singularity, superiority, Thursday, today, truth, Tuesday, uniformity, universality, university, Wednesday, yellow.

Two examples of the set of words associated with each concept are shown hereafter. The examples are given in the languages used in this research project. They are conventionally annotated using the cursive form for words which are not correctly spelled, but which will be measured anyway (like French words after removing diacritics), and UPPER CASE for words which suffer from cross-linguistic homography or other complications (and then may deserve a special processing).

22 Note that the counting of word by search engines is independent of upper or lower cases.

Table 4: Example of word-concepts

English	Spanish	French	Italian	Portuguese	Romanian	German	Catalan
fidelity fidelities faithfulness faithfulnesses	fidelidad FIDELIDADES	fidélité fidelite fidélités fidelites	Fedeltà fedelta	Fidelidade FIDELIDADES	fidelitatie fidelitatea fidelității fidelității fidelități fidelitati fidelitățile fidelitatiile fidelităților fidelitatiilor	TREUE TREUEN	fidelit FIDELITATS
Monday Mondays	Lunes	lundi lundis	lunedì lunedì	segunda- feira segundas- feiras	luni lunea	montag MONTAGES montags MONTAGE MONTAGEN	Dilluns

The full table, which holds just over 1700 words, can be consulted online at: <http://funredes.org/lc/english/historia/listapa.htm>.

How was this final list of word-concepts obtained? First, by establishing a set of criteria to obtain the best word-concepts with multilingual equivalents. From there, a large number of potential word-concepts were tested and filtered²³. However, it was impossible to obtain perfect results and some post-processing was duly required to avoid statistical bias (such as the split of the number of citations of *fidelidades* between Spanish and Portuguese, or the count for *montage* and *montages* to be deduced by separating the number of occurrences in French from those in German).

In some cases, a recognized problem was considered as acceptable because only a marginal impact on the statistical process was expected (like in the example the fact that German TREUE and TREUEN are also form of the adjective “faithful” thus giving additional semantic to the German words).

Below is the complete list of criteria used to create the sample of word-concepts.

²³ As a matter of fact various hundreds of words were examined prior to reach the final table and the process was the result of an intense team work lasting several months and marked by the intensity of collaborative exchanges between Union Latine and FUNREDES teams and within each institutions.

Criteria 1: Cultural neutrality

Definition: property of a word in relation to the frequency of its appearance in a given language, considering its cultural value and meaning.

Examples: *Wine*, *perfume*, *gastronomy* are not culturally neutral in French.

Rule: reject terms which are obviously culturally sensitive.

Criteria 2: Cross-language homography

Definition: the orthography of a term in one language is identical to a term in another language, regardless of whether the meaning is the same or not. Strong homography occurs when the orthography is the same, including diacritics. Weak homography occurs when the only difference between terms is due to diacritics.

Examples: *casa* has the same meaning and writing in Spanish and Portuguese; *red* means *network* in Spanish; *hier* means *yesterday* in French and *here* in German. Homography can also be found in one part of a composed word like in *mardi-gras*, although *mardi* means *Tuesday* in French.

Rules:

- avoid concepts which include words of less than four letters, so as to reduce the probability of homographies with languages out of the scope of the study;
- when a word with homography is present in the sample (this is quite unavoidable for Spanish and Portuguese), then split the number of pages in proportion to the relative presence of each language (the words marked in upper case in the table are subject to this rule);
- whenever possible, correct the computed numeric results by removing the count of the homographic word or expression (for instance, the score of *mardi* in French is obtained after subtraction of the count of *mardi-gras* in English).

Criteria 3: Homography by acceptance

Definition: when a word in a given language is also used in other languages.

Examples: The English words *business*, *sandwich* or *software* are used in many other languages in their English form. The French's *déjà vu* is used in English.

Rule: Discard the concepts which include such words.

Criteria 4: Homography with abbreviation

Definition: when a word in a given language has the same orthography as that of an abbreviation frequently used by other languages.

Example: The number *sept* (*seven* in French) is homographic of the abbreviation of *September* in English.

Rule: Discard the concepts which include such words. Note that the rule to avoid words of less than four letters reduces the probability of such occurrences.

Criteria 5: Homography with a frequent given name

Definition: when a word in a given language has the same orthography as that of a common given name.

Example: *Julio* means *July* in Spanish, but it is also an extremely frequent first name in Spanish. *Windows* means part of a building but it is also the name of a software trade mark frequently quoted in the Internet.

Rule: Discard the concepts which include such words.

Criteria 6: Loose homography

Definition: when the orthography of a word with a single spelling mistake corresponds with an existing word.

Example: *Embassador* in English if written with only one "s" correspond to the same concept in Romanian.

Rule: Discard the concepts which include such words if and only if the target language is English (the only case where it could have a significant statistical impact).

Criteria 7: Existence of non-equivalent meanings for the same word

Definition: when the same word has different meanings which are expressed by different words in other languages.

Example: *Prix* in French means both *price* and *prime (or award)*.

Rule: Discard the concepts holding such words or ensure the inclusion of all the corresponding meanings in every language.

Criteria 8: Non equivalent morpho-syntax (verb, noun)

Definition: when the same word has different syntactic meanings (i.e. it corresponds to both a verb and a noun) which are expressed by different words in other languages.

Example: *Love* in English (both a noun and a verb) corresponds to French noun *amour* and verb *aimer* in different conjugations (*aime, aimes, aimons, aimez, aiment...*).

Rule: Avoid such words. Note that this is the reason why the sample does not include any verbs.

Criteria 9: Non-equivalent morpho-syntax (adjective, noun)

Definition: when the same word has different syntactic meanings (corresponding to adjective or nouns) which are expressed by different forms of words in other languages.

Examples: *Yellow* corresponds to Spanish's *amarillo, amarilla, amarillos, amarillas*. The couplet (*instability, instabilities*) corresponds to the following Romanian list: *instabilitate, / instabilitatea, / instabilități / instabilități / instabilitățile / instabilităților*.

Rule: These words are acceptable provided a careful screening is undertaken to ensure syntactic equivalence. This explains why the number of words in the sample can vary depending on the language.

Criteria 10: Lexical-semantic diversity

Definition: when the same concept is expressed by different words depending on the country using the same language.

Examples: Depending on Spanish-speaking country in question, the English word *gasoline* is said as *nafta* or *gasolina* in Spanish.

Rule: These words are acceptable for inclusion providing a careful screening is undertaken to allow for syntactic and semantic equivalence in the different countries.

Criteria 11: Orthographic diversity

Definition: when a given word has diverse spelling depending on the country.

Examples: *Theater* in the United States and *theatre* in United Kingdom. *Electricidade* in Portugal and *eletricidade* in Brazil (without “c” before “t”).

Rule: These words are acceptable providing a careful screening of the different spellings is undertaken.

The application of these eleven criteria allowed the actual list of word-concepts to be built and left unchanged during the whole process. Additional languages were measured and added to the table, included as a new column, listing words corresponding to the language in question.

Post-processing

As shown in the two examples tabled above, complete filtering of the word-concepts to eliminate all linguistic problems was not possible (as the probability of existing cross-linguistic homography is quite high). Some post-processing was required in order to reduce the statistically-unwanted biases. The decision about whether or not

to correct the figures obtained was based on pragmatic, statistical considerations. A correction was generally made using a simple rule of linguistic proportionality, following an assessment of the apparent prevalence of the considered language on the Internet. The most common situation was the plural form, such as the form of *idades*, which is quite common in Spanish and Portuguese. It was decided to split the number of such occurrences between the languages in question, according to their proportionate, computed, online presence. There are several other situations which have required some processing, which are all exposed in previous reports²⁴.

All post-processing has been integrated into the project programming and completed without human intervention. However, for each unit of measurement, a complete manual screening of the search engine results was systematically conducted (and aided by the computer program which highlights statistical anomalies). This helped detect possible conflicting situations, which could occur due to homographic abbreviation, or certain given names gaining a new, strong presence on the Internet.

Attempt to change the sample

At one stage, an intensive trial was undertaken using expressions²⁵ composed of a few words instead of single word-concepts, in order to overcome the probability of homography. However the time investment was not fruitful, as the results obtained were erratic. In many cases, the count for the occurrence of the English expressions was significantly less than that of the other languages being studied. An analysis of this phenomenon led the team to attribute these chaotic behaviours to the loss of linearity of the embedded mathematical function.

The following examples illustrate what happened:

24 See for instance <http://funredes.org/lc2005/english/L4index.html>

25 A new sample of more than 200 expressions was created.

In English	Number of occurrences	In French:	Number of occurrences
“networks”	3,834,260	“reseaux”	326,250
“development”	21,258,510	“developpement”	909,790
“networks and development”	201	“reseaux et developpement”	61
“note bank” = 150,000, “billet de banque” = 128,000 , “billete de banco” = 18,700			

Thereafter the team decided to keep the sample. The sample had demonstrated its validity by providing coherent results over time, and demonstrated comprehensible changing trends in accordance with search engine evolution.

4.2 SEARCH ENGINE METHODOLOGY

The whole process of this study has been characterized by a permanent struggle with search engine behaviours. The primary activity for each measurement campaign was to validate whether the search engines could meet the methodological requirements of the research, and, in many cases, understand the rationale for what first appeared as invalid results. The establishment of the linguistic methodology was a once-off yet important investment for this project. Unfortunately, the changing nature of search engine indexes and their functionality transformed this part of the work into the most unpredictable way.

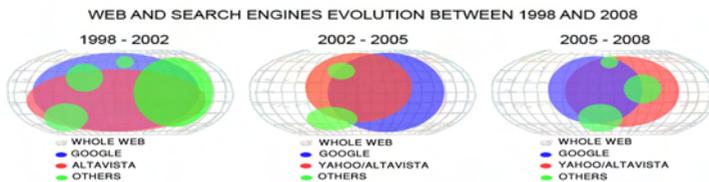
The method implied careful checking of how every available and independent search engine behaves, in terms of diacritics and page counts. In the context of a lack of transparency (and frequent changes) on behalf of the search engine providers, the tests were multiplied many times in order to understand some of the erratic results yielded. On several occasions, these results precipitated a high decrease of trust in the method until a prolonged, collaborative effort helped the researchers discover the reason for the strange data produced. This demonstrated that the results produced by the search engines were not completely reliable²⁶. The corresponding

²⁶ This is unfortunately the situation with Google since 2005, which has forced the team to discard it for the study after months of trials, and instead use Yahoo (Altavista). The number of occurrences displayed by Google for a certain word, setting language and domain parameters to “any” being, counter-logically, much lower than the sum of the number of occurrences of the same word by language or by domain...

part of the project methodology was then fixed accordingly using computer programming.

On a positive note, during the first series of measurements undertaken using the fully-defined methodology, a satisfactory outcome was achieved. All the search engines used at that time yielded results that not only were statistically validated, but also extremely close to each other. This strengthened the trust in the methodology employed. However, over time, the situation changed. The following diagram explains why the situation today is fraught with challenge.

Figure 1: Search engines coverage over time



Nowadays, there are three phenomena which indicate the absurdity of the continuation of the original methodology that was based on the use of search engines:

- 1) Search engine indexes now represent less than 30% of the total cyberspace universe (compared to more than 80% in the past) and are more and more responsive to hidden commercial criteria which considerably increase the linguistic bias towards that of English language predominance²⁷.
- 2) Search engines are becoming more “intelligent” (for instance they will search for concepts in different languages), thus rendering the methodology of simple page counting senseless;
- 3) The rise of advertisements embedded in Web pages is bringing new biases to the research results²⁸.

It could still be argued that the method used enables a comparison of the linguistic biases of the different search engines. However, this would not make it any more possible to pretend that the results for

27 This may open durable loss of niche for Google as in the case of using Exalead to search the French speaking Web.

28 It is more and more frequent that non English pages are completed with dynamic advertisements in English.

a given search engine could be extrapolated to provide an accurate representation of the proportionality of diverse languages for the whole Web.

As a matter of fact, the graphical representation of the evolution of languages positions following the measurements (see Chapter V Results) has already been altered by important changes in search engine behaviour. In 2001, the online presence of all the languages measured in the study was decreasing proportionately in comparison with English, and in contradiction with the observed trend. Was it that English was suddenly bouncing back on the Web? Was it that the growing online Asian presence was boosting English proliferation? Careful and patient analysis²⁹ led the team to conclude that this situation was merely a reflection of the reshaping of the Google index which, in a transition phase, was increasing its bias towards English (a bias which has always existed in some proportion anyway).

Between 2003 and 2004, Google and Yahoo were the two best search engine options in terms of meeting the requirements for linguistic measurement online. The large size of the Google index (three billion pages) and its clear management of diacritics at that time led the team to choose Google as the primary search engine. MSN was discarded for having a strong bias towards English, and similarly, Exalead was identified as having a bias towards French. Most of the other search engines merged with those mentioned above, or had too small an index to be useful for the purposes.

In 2006, the study faced a prolonged period of incoherent results for four measurements and it proved impossible to find a rationale in the search engines' behaviour. As a matter of fact, the project was very close to being terminated due to these results. However, an explanation was finally found in Google's so-called "*Big Daddy*³⁰ operation", which consisted in a total reshaping of its index and the servers hosting the data base. This redefinition and rebuilding of the index necessitated a long transition time. It appeared obvious that the rebuild had a clear tendency to begin with the English Web

29 Months were spent checking and discarding different hypothesis to explain such situation, like for instance if the burst of Asian countries was triggering a surge of English.

30 <http://www.webworkshop.net/googles-big-daddy-update.html> or <http://www.mattcutts.com/blog/bigdaddy>

before other languages. This totally disturbed the results during this transition phase. Progressively, the results began to once again yield figures consistent with those from previous measurement units³¹ and confidence was regained, only to conclude several months later that Google was undertaking more changes which definitely made it unusable for the project. This last fact was only realized after several months of work in 2007. It obliged to return to Yahoo (which uses the search engine of Altavista), until it was finally decided to find another manner to pursue the quest (which will be explained later in the paper).

The following sections provide more detail about specifics of the methodology employed, including statistical analysis, process and results.

4.3 STATISTICAL METHODOLOGY

The set of 57 values of the total of number of pages that counts each word-concept in each language, divided by the respective value of the same world-concept in English (representing the percentage of a given language compared to English), was processed as a *statistic random variable* for which the traditional tools for a *Gaussian function* (or *normal distribution*) were applied. The *coefficient of variance*³² was then computed. A value of 0 would indicate a constant result (which is an absolutely impossible result); whereas a value of 1 would indicate an exponential function, representative of a *normal* random situation. Between 0 and 1 the coefficient of variance would indicate a good result (with low deviation). A value superior to 1 would then precipitate questions about the validity of the method, indicating a hyper-exponential function representing excessive dispersion. This value was used to control the measurements, as attention was given for results above 1 and, in general, some anomaly in the process could be detected. Generally speaking, the measurements always offered credible statistical results based on that indicator. Then the

31 The best warranty of the method is the fact that new results of measurements always show some kind of continuity with the historical results and made an imperative point to support change of trends in the results by some sound arguments about what was happening in the field (like for instance when Spanish slide below French after the first surge of Internet users in Latin America back in 1999).

32 The squared root of the squared standard deviation divided by the squared average.

confidence interval for 90% and 99% was computed using the Student-Fisher law and allowing the validity of the results within a window to be located.

4.4 INDICATORS BUILDING

The first indicator built was related to the presence of a given language on the Internet relative to its presence in the real world (or *weighted presence*). A ratio of 1 expresses normality, whereas a ratio of below 1 expresses a weak virtual presence (as it was found for Spanish and Portuguese in the first editions) and a ratio of above 1 indicates a strong virtual presence (which is obviously the case for English and, to lesser extent, for French, Italian and German). The evolution of this indicator for a given language demonstrates how it could enhance its virtual presence and reach a normal ('real world') presence, or even higher. In the case of English, which still has a value significantly higher than 1, the 12 years of measurement have shown a steady decline in its value (from 7 to 4) and then the stabilization of its position (with the previously-explained limitation that the search engine index results after 2005 cannot be extrapolated to the whole worldwide Web). This indicator could be useful to measure the efficiency of a virtual linguistic policy.

Using estimates of the number of Internet users for a given language (which were provided for many years by GlobalStat³³ and since 2005 by Internet Worldstats³⁴), it is possible to build an indicator of *linguistic productivity* (the number of pages produced by Internet users which is normalized to have 1 as the average). However, one should be warned about the limits of the reliability of the figures of those organizations. The research team estimated the figures provided are correct only within a margin of more or less 20%, since the methodology is based on data provided by multiple national sources which may not have standardized approaches. This obviously impacts the figures produced by FUNREDES/Union Latine in the same proportion.

One of the first interesting results of the measurements was to discover that the gap between high and low productivity languages was not so important. Most of the measured languages were close

33 <http://global-reach.biz/globstats/index.php3>

34 <http://http://www.internetworldstats.com/>

to 1. This implies some kind of natural rule between the proportion of content producers and the total number of Internet users. Translated in terms of linguistic policy, this indicates that the most obvious policy needed to boost content in a given language is firstly to increase the number of Internet users. Another interesting lesson learnt from this indicator was that the apparently natural law of proportionality tended to lose clout in recent years. This could be interpreted by the fact that Internet users who were late adopters of Internet technology, tend to be content consumers rather than content producers (despite the boom of blogs). This fact supports the argument for new policies more oriented towards digital and information literacy than mere Internet access.

Some additional indicators for each language are shown in the following table (2007 figures) and provide interesting content for further analysis:

Table 5: Indicators for languages in the Internet (2007)

	EN	SP	FR	IT	PO	RO	GE	CAT	Total
Speakers (millions) ³	670	400	130	60	205	30	120	9	6607 ⁴
Speakers as % of world population	10.1%	6.1%	2.0%	0.9%	3.1%	0.5%	1.8%	0.1%	130% ⁵
Internet users in a given language (millions) ⁶	366	102	58	31	47	5	59	2	1154 ⁷
Internet users in % of speakers	54.6%	25.4%	44.9%	52.3%	23.1%	16.5%	49.1%	23.1%	17.5% ⁸
Internet users in % of world population	5.5%	1.5%	0.9%	0.5%	0.7%	0.1%	0.9%	0.0%	17.5%
% internet users per language	32%	9%	5%	3%	4%	0%	5%	0.2%	130%
% web per language ⁹	45.0%	3.8%	4.4%	2.7%	1.4%	0.3%	5.9%	0.1%	100%
Ling. productivity per language ¹⁰	1.42	0.43	0.87	0.98	0.34	0.66	1.16	0.74	1
web pages per internet users in a given language	4.44	0.63	2.24	2.93	0.45	0.62	3.25	0.96	

The ratio between speakers and Internet users in a given language (*Internet users as a % of speakers*) is one type of indicator of language penetration on the Internet. It informs the future evolution of its potential growth. For example, when a language reaches 50% penetration, it is to be expected that the inflexion on the curve of growth has been reached, and the remaining growth will start to be asymptotic. The main reason for the relative decline of English on the Internet is therefore simply because it has already peaked, by reaching an early and transitory huge, initial presence³⁵.

The *percentage of Internet users per language* as a proportion of the total user population is another indicator of the linguistic and digital divide. For example, it could be explained as follows: “17.5% of the world population was connected to the Internet in 2007, of which 5.5 % were English speaking Internet users”. The *percentage of Internet users per language* is also important to understand. It helps reveal the linguistic diversity on the Internet and has shown strong and steady changes since the beginning of the World Wide Web.

The second set of indicators, oriented to a measurement that focuses more on individual countries (and requires new methodological tricks) has been produced since 2001. They present extremely rich information about the dynamics of content production by language and by country. These indicators have been progressively generalized in the study to include French, Spanish, English and Portuguese (four languages which are used in many different countries and for which it is interesting to observe and compare the contribution of each of these countries).

This has been achieved thanks to search engine capabilities that enable the team to measure the number of occurrences of pages mentioned for a given search by country. The program was run multiple times for different countries, to obtain the results presented.

Here the methodological difficulty is that it is insufficient to measure the sample by country code top level domain name (ccTLD), since many servers for any given country use generic top level domain

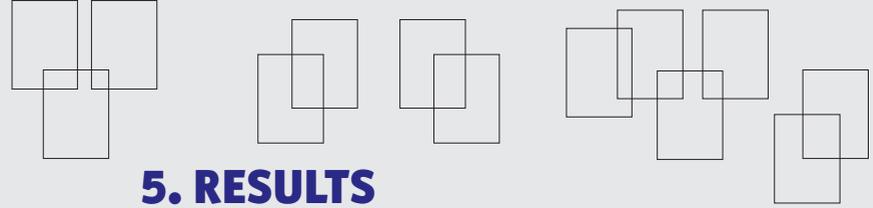
³⁵ The experience of the massive experience of Minitel in France shows that when reaching 60% of penetration the remaining growth remains very slow in spite of the absence of direct cost.

names (gTLD)³⁶. This means that the contents of the gTLD must then be split between the countries. This is achieved by estimating the percentage of domain names using the ccTLD. These figures are obtained directly from colleagues working in Network Information Centres or by literature screening.

The results are extremely powerful, and are excellent tools for policy makers, as they can be compiled by region, providing a measured indication of the digital divide between the South and the North. Furthermore, they can also provide an idea of what foreign language contents are produced by some countries. Yet some caution should be exercised and the results should only be taken as approximate, as the methodology does not allow very precise data³⁷. More details on the results are given in chapter 5.3.

36 Mainly the .com and .org gTLDs.

37 Especially in relation with the percentage of US Web site which are actually under the .US TLD, a figure which is hard to know and that has been computed by trial and error to reach a 100% in the total.



5. RESULTS

5.1 MAIN RESULTS

The main results from the series of measurements are presented below, in both graphical and table format.

WEB PRESENCE PERCENTAGES COMPARED TO ENGLISH

The following table quantifies comparative linguistic presence in cyberspace. Expressed as percentages in comparison to English (where English is 100%), the table reads as follows: as of September 1998, for 100 Web pages in English, there were three pages in Spanish, four pages in French, two pages in Italian and one page in Portuguese. In order to have one page in Romanian, 500 pages in English were needed.

Table 6: Web presence of studied languages compared to English

	SP	FR	IT	PO	RO	GE	CAT
09/98	3.37%	3.75%	2.00%	1.09%	0.20%		
08/00	8.41%	7.33%	4.60%	3.95%	0.37%	11.00%	
01/01	9.46%	7.89%	4.93%	4.44%	0.33%	11.43%	
10/01	11.36%	9.14%	6.15%	5.61%	0.36%	14.08%	
02/02	11.60%	9.60%	6.51%	5.62%	0.33%	14.41%	
02/03	10.83%	8.82%	5.28%	4.55%	0.23%	13.87%	
02/04	10.30%	10.18%	6.09%	4.36%	0.41%	15.35%	
03/05	10.23%	11.00%	6.77%	4.15%	0.37%	15.42%	
11/07	8.45	9.80%	5.92%	3.09%	0.63%	13.12%	0.30%

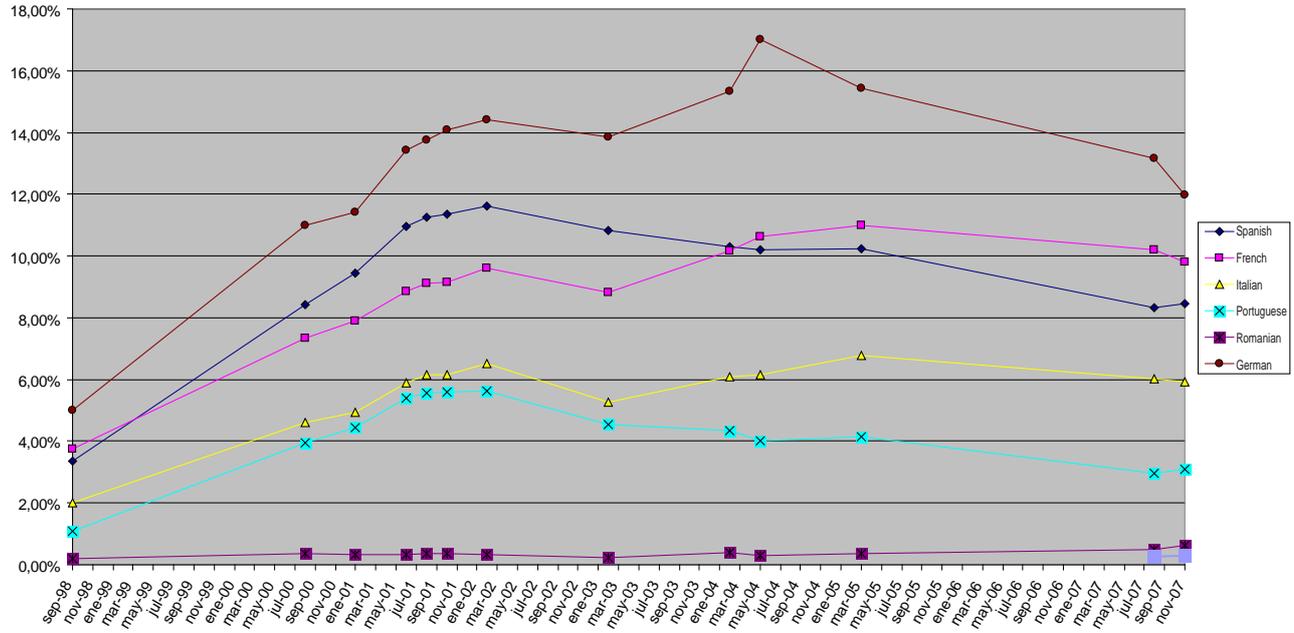
Table 7: Confidence intervals for 2008 results

	99%	90%	90%	99%
Spanish	6.56	7.48	10.74	11.66
French	8.19	8.98	11.77	12.56
Portuguese	4.01	4.65	6.9	7.53
Italian	1.82	2.24	3.73	4.15
Romanian	0.52	0.63	0.99	1.09
German	7.78	8.53	11.18	11.93
Catalan	0.25	0.29	0.44	0.49

This above table can be read as follows: there is 99% probability that the percentage of French Web pages compared to English is between 8.19% and 12.56%. There is 90% probability that the percentage of Italian Web pages compared to English is between 2.24% and 3.73%.

Figure 2: Graph of evolution of studied languages' percentages compared to English

Evolution of Latin languages compared to English



The analysis of the graphs above shows two phenomena which have been described in the methodology section of this paper:

- In 2003, all the languages measured in terms of their online presence declined in the same proportion as compared to English. This was finally interpreted as a transitory situation, due to a change in Google's indexing rather than the declining presence of those languages on the Web. Extrapolating the results from 2002 to 2004 would probably have provided a fairer depiction of reality for that period, as the graph tended to indicate.

- Starting in 2005, a parallel decline of the measured languages was also visible, as depicted in the graph. Thereafter, it was unfortunately impossible to extrapolate the results of the search engines' indexes as a fair representation of virtual reality on the Web. Rather, what is measured has to be read as just the reality within the Web pages indexed by a specific search engine. This indicates a new growing bias in favour of English for the most generic of the search engines.

As for the studied languages, what is noticeable is the initial push for increased Internet access for Spanish and Portuguese language speakers, driven by Latin America between 1998 and 2002. This was followed by the relative weakening of Spanish and Portuguese online presence compared to French, German or Italian. A strengthening of the Romanian presence on the Web started much later, in 2007, and its development should continue to be monitored to see if it will also plateau.

The following table provides an estimate of the absolute presence of languages on the Web. It was obtained by making an estimate of English and then applying the comparative percentages for other languages from the study. The estimation for English is made by iteration, playing with the value of the rest of languages. It is increasingly more difficult to make this estimate with confidence, due to the explosion of users in Asia and also taking into account search engine bias (towards English).

Table 8: Absolute percentage of studied languages in the Web

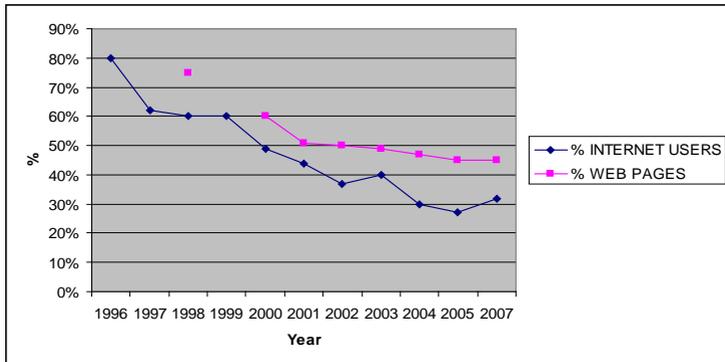
	EN	SP	FR	IT	PO	RO	GE	CAT	SUM ¹¹	REST ¹²
09/98	75.0%	2.53%	2.81%	1.50%	0.82%	0.15%	3.75%		11.56%	13.44%
08/00	60.0%	5.05%	4.40%	2.76%	2.37%	0.22%	3.00%		17.80%	22.20%
01/01	55.0%	5.20%	4.34%	2.71%	2.44%	0.18%	6.29%		21.16%	23.84%
06/01	52.0%	5.69%	4.61%	3.06%	2.81%	0.17%	6.98%		23.31%	24.69%
08/01	51.0%	5.73%	4.66%	3.14%	2.84%	0.18%	7.01%		23.55%	25.45%
10/01	50.7%	5.76%	4.63%	3.12%	2.84%	0.18%	7.14%		23.68%	25.62%
02/02	50.0%	5.80%	4.80%	3.26%	2.81%	0.17%	7.21%		24.04%	25.97%
02/03	49.0%	5.31%	4.32%	2.59%	2.23%	0.11%	6.80%		21.35%	29.65%
02/04	47.0%	4.84%	4.78%	2.86%	2.05%	0.19%	7.21%		21.94%	31.06%
05/04	46.3%	4.72%	4.93%	2.85%	1.86%	0.14%	7.88%		22.38%	31.32%
03/05	45.0%	4.60%	4.95%	3.05%	1.87%	0.17%	6.94%		21.57%	33.43%
08/07	(45.0%)	3.75%	4.59%	2.70%	1.34%	0.23%	5.93%	0.12%	18.53%	
11/07	(45.0%)	3.80%	4.41%	2.66%	1.39%	0.28%	5.90%	0.14%	18.46%	

The apparent asymptotic curving of English towards 45% (see Figure 2) is due to the new bias of search engines, rather than a real phenomenon of the linguistic topology of the Web. If the curve of English speaking users is a fair indicator of trends, as it should be (see below), then the English presence on the Web (as opposed to its presence through search engine indexes) is probably below 40%; the last column values for 2007 suffer from the same problem. The reality is probably above 40% for the rest of the languages, due mainly to a massive Chinese online presence.

Table 9: Evolution of percentages of English speaking Internet users and web pages

	96	97	98	99	00	01	02	03	04	05	07
Internet users¹³	40	72	91	148	192	231	234	288	280	300	366
% internet users	80%	62%	60%	60%	49%	44%	37%	40%	30%	27%	32%
% web pages			75.0%		60.0%	51.0%	50.0%	49.0%	47.0%	45.0%	45.0%

Figure 3: Evolution of percentages of English speaking Internet users and web pages (graph)



The sharp increase in Internet user percentages in English indicated in the graph above, between 2005 and 2007, is the result of the change of source from GlobalStat (which has stopped providing such statistical information) to InternetWorldStats. It is a consequence of the limitations of those figures discussed in chapter 4.4.

5.2 ANALYSIS PER COUNTRY

The most interesting and innovative results of the study were obtained thanks to the application of the method that used domain names for the English, French, Spanish and Portuguese speaking countries. This approach produced extremely striking data.

The full results for each language can be found at the following Websites:

- <http://funredes.org/lc/english/medidas/sintesis.htm>
(for 2005 results)
- http://dtiil.unilat.org/LI/2007/index_es.htm
(for 2007 results)

The number of interesting results is too high to be described in detail in this paper. A synthesis of the results produced is provided below:

Table 10: Main countries producing Web pages in French: percentage of pages followed by productivity³⁸

	11/2007	5/2005	3/2003
FRANCE	60% -1.09	60% - 0.82	54% - 0.96
CANADA	20% - 1.06	19% - 1.27	24% - 1.83
BELGIUM	7% - 0.60	8% - 1.55	7 % - 2.21
SWITZERLAND	5% - 0.87	5% - 2.78	6% - 2.17
OTHERS	8% - 0.84	8% - 1.38	9% - 3.10

Canada (and especially Québec) was one of the earliest content producers on the Web and this is why it appears to be decreasing in terms of productivity over time. Conversely, France was a relative latecomer, booming in 2005.

Two trends to be noticed from this table: first, a general decrease in productivity (except for France) and second, a strong decrease in productivity for Belgium and Switzerland, due to a lot of new Internet users and not much new content or web page production.

Table 11: Web pages in French: production by region

	11/2007	5/2005	3/2003
EUROPE	75%	79%	71%
AMERICA	22%	21%	25%
AFRICA/ARAB STATES	0.3%	0.4%	0.4%
ASIA/PACIFIC	0.2%	0.4%	0.4%
NOT CLASSIFIED	2.11%	0.19%	3.32%

The sad reality of the digital divide is obvious in the above table, looking at the result for Africa. To date, no change has been noted for this trend. The developing Francophone countries which appear the most productive in 2007 are Morocco and Senegal, although it is worth noting that Germany or the United Kingdom (UK) actually produces more French pages than all African countries combined.

As for the Spanish language, the following table shows the main content producers and their associated productivity:

³⁸ Computed as the ratio of % of production per % of Internet user in the given language.

Table 12: Main countries producing Web pages in Spanish: percentage of pages followed by productivity

	2007	2005	2001
SPAIN	56% - 3.4	48 % - 2.4	54% - 2.7
USA	10% - 0.4	14% - 0.4	5 % - 0.12
ARGENTINA	9.4 % - 0.9 ¹⁴	10.6% - 1.9	9.6% - 1.3
MEXICO	8.4% - 0.45	7.4 % - 0.5	8.6 % - 0.45

In 2001, the United States of America (USA) had more Spanish speaking Internet users than Spain. Yet Spain produced 54% of the total of Spanish content online and the USA produced only 5%. Since then, online productivity for the USA has improved, but did not reach the average factor of 1. It is noticeable that Mexico, which has the highest population of migrants in the USA, has the same low figure for productivity of Spanish Web pages. This can be interpreted in terms of public policy for Spanish content creation, to focus the virtual borders between the USA and Mexico.

The highest productivity ratio was found in Cuba, which increased from 3.4 in 2001 to 4.3 in 2007. Notably, Nicaragua is not far behind. This demonstrates both the low number of Internet users and a policy of systematically publishing on the Web by academia in these countries.

Table 13: Main countries producing Web pages in English: percentage of pages followed by productivity

	11/2007	5/2005
USA	66% - 1	51 % - 0.8
UK	6.5% - 0.6	7.2 % - 0.6
CANADA	3.5% - 0.7	5 % - 0.7
AUSTRALIA	1.5 % - 0.3	1.8 % - 0.4
GERMANY	1.2 % - 39	1.9 % - 57

Germany's inclusion in the table above, with a quantifiably large output of Web pages, illustrates a phenomenon which was predictable: many countries for whom English is not the main language also notably contribute to the production of Web content in English.

Some ccTLDs of very small island states show abnormally high results for content production in English. This is the result of their ccTLD being sold for foreign commercial purpose (like Tuvalu with .tv, Niue with .nu, Micronesia with .fm and Samoa with .ws).

Consideration of linguistic diversity online and the number of Internet users and page output by both country and region demonstrates the extent of the digital divide. The overall results show that the entire output of Web pages produced by African countries in English or French hardly reach 0.33% of the total number of Web pages produced globally for those languages. Of the statistics for Africa, 97% of content generated is by South Africa. Other non-English-speaking Organisation for Economic Co-operation and Development (OECD) countries produce more than 0.1% of this total, which is a third of the whole production of Africa. Furthermore, many individual countries, like Germany, France, Italy or Japan, produce more English content than all African countries combined, including South Africa.

Together with indicators built by the Language Observatory Project (LOP)³⁹, this situation projects a message that ICT4D groups still resist integrating in their plans: the digital divide is as much if not more about content production, as it is about access to the Internet (see [Accessing content](#) in references). The content divide, which is a linguistic and cultural divide, is a worrying indicator of the risk of acculturation of populations that gain access and have no choice of content in their mother tongue. This should drive a rebalance of the digital divide policies and give much more priority to digital and information literacy (which are obvious triggers of content production and information ethics). The struggle against the digital divide is not a mere question of access and infrastructure.

Table 14: Main countries producing Web pages in Portuguese: percentage of pages followed by productivity

	11/2007	5/2005
BRAZIL	71% - 0.90	71 % - 0.95
PORTUGAL	15 % - 0.98	17 % - 1.0
USA	4 % - 5.0	8 % - 5.4
SPAIN	3.8 % - 3.7	2.3 % - 1.2

39 <http://www.language-observatory.org/>

Table 14 above indicates that Brazil dominates the production of Web pages in Portuguese. It has retained a stable content production, producing 71% of all Web pages that appear on the Internet in Portuguese. It is also worth noting here that the USA produces more Web pages in Portuguese than Spain.

5.3 OTHER SPACES FOR LANGUAGE DIVERSITY

Linguistic diversity can be studied in a number of other areas of cyberspace. In the first years, measurements were made in the Usenet space (Newsgroups) with interesting results⁴⁰. More recently, Blogs have also been examined for their linguistic diversity. The results for different Blogs' search engines are so heterogeneous that they are not worth publishing. The fact is that today, each Blog's search engine is associated with a specific server that only searches the index for that particular Blog. In the future, a "meta-search engine" adding up the results of the various Blogs search engines could be a useful service indeed.

More interesting are the results obtained by studying Wikipedia, which maintains fascinating multilingual statistics⁴¹, confirming that it is one of the most linguistically-diverse spaces of the Internet. An analysis reveals the number of articles that appear in a number of given languages, as tabled below (Source: Wikipedia, July 2008):

Table 15: Wikipedia articles per language

English	2.259.431	23.078%
German	715.830	7.312%
French	629.004	6.425%
Polish	475.566	4.857%
Japanese	472.691	4.828%
Italian	418.969	4.279%
Dutch	413.325	4.222%
Portuguese	363.323	3.711%
Spanish	337.860	3.451%

40 <http://www.funredes.org/lc2005/english/L4index.html>

41 http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

5.4 CULTURAL DIVERSITY

The methodology used to gauge the presence and distribution of cultural diversity on the Internet is quite simple, perhaps even simplistic. It can only be taken as a first, basic approximation. It does not really reflect the complexity of the subject in question, which can be measured on a thematic as well as 'national' basis, amongst other ways. To provide an indication of cultural diversity, a number of themes were selected and for each of them a large, but far from exhaustive, list of pertinent personalities has been provided (such as Albert Einstein for Science or Pablo Picasso for Graphic Arts). The 'Web citation index' for each personality was computed and the results were compiled. A simple indicator was then devised and used to track the evolution of cultural diversity online over a series of measurements undertaken in 1996, 1998, 2001, 2005 and 2008. Considered comprehensively, these provide a perspective of cultural diversity over the last 12 years.

The themes which have been considered as appropriate to the measurement of cultural diversity are:

- Literature
- Science
- Music (all types)
- Cinema
- Graphic arts
- Politics
- People (persons who are famous and present in media for any reason)
- History
- Fiction (such as Dracula or Cinderella)
- One word (persons from any theme in one word, like Einstein or Picasso).

A total of some 1200 personalities⁴² was computed and some examples of the results are listed below (the complete results can be consulted in the Web⁴³). Note the results are colour coded to

42 A change was made in the second measurement to reach a more complete sample of personalities. After that, the same sample was kept.

43 <http://funredes.org/lc/espanol/cultura08/cultura08.htm>

help identify their cultural categorization⁴⁴ and the second column indicates the change in the hit parade compared to the previous measurements).

Table 16: First positions in literature

2008			2005			2001		
1	WILLIAM SHAKESPEARE	0	1	WILLIAM SHAKESPEARE	0	1	WILLIAM SHAKESPEARE	0
2	OSCAR WILDE	2	2	RENÉ DESCARTES	26	2	VICTOR HUGO	1
3	VICTOR HUGO	3	3	GABRIEL GARCÍA MÁRQUEZ	34	3	OSCAR WILDE	-1
4	CHARLES DICKENS	4	4	OSCAR WILDE	-1	4	CHARLES DICKENS	2
5	AGATHA CHRISTIE	21	5	J.R.R. TOLKIEN	7	5	WILLIAM JAMES	0
6	PAULO COELHO	3	6	VICTOR HUGO	-4	6	JAMES JOYCE	2
7	J.R.R. TOLKIEN	-2	7	LORD BYRON	14	7	ERNEST HEMINGWAY	7
8	ERNEST HEMINGWAY	15	8	CHARLES DICKENS	-4	8	WALT WHITMAN	-1
9	EDGAR POE	9	9	PAULO COELHO	53	9	EDGAR POE	-5
10	JULES VERNE	1	10	IMMANUEL KANT	20	10	HENRY JAMES	1

Table 17: First positions in science

2008			2005			2001		
1	ALBERT EINSTEIN	0	1	ALBERT EINSTEIN	0	1	ALBERT EINSTEIN	0
2	NOAM CHOMSKY	1	2	MARIE CURIE	0	2	MARIE CURIE	9
3	CHARLES DARWIN	1	3	NOAM CHOMSKY	4	3	CHARLES DARWIN	0
4	MARIE CURIE	-2	4	CHARLES DARWIN	-1	4	SIGMUND FREUD	0
5	SIGMUND FREUD	4	5	ISAAC NEWTON	0	5	ISAAC NEWTON	-3
6	ISAAC NEWTON	-1	6	BLAISE PASCAL	4	6	THOMAS EDISON	0
7	THOMAS EDISON	5	7	GALILEO GALILEI	4	7	NOAM CHOMSKY	0
8	CARL SAGAN	2	8	ALEXANDER VON HUMBOLDT	4	8	LOUIS PASTEUR	0
9	MILTON FRIEDMAN	4	9	SIGMUND FREUD	-5	9	CARL SAGAN	-4
10	GALILEO GALILEI	-3	10	CARL SAGAN	-1	10	BLAISE PASCAL	-1
11	BLAISE PASCAL	-5	11	LOUIS PASTEUR	-3	11	GALILEO GALILEI	-1
12	LOUIS PASTEUR	-1	12	THOMAS EDISON	-6	12	ALEXANDER VON HUMBOLDT	2

⁴⁴ American English has been differentiated from the European (UK-based) English.

Table 18: First positions in graphic arts

2008			2005			2001		
1	LEONARDO DA VINCI	0	1	LEONARDO DA VINCI	0		LEONARDO DA VINCI	0
2	ANDY WARHOL	1	2	SALVADOR DALÍ	1	2	ANDY WARHOL	0
3	SALVADOR DALI	-1	3	ANDY WARHOL	-1	3	SALVADOR DALÍ	0
4	PABLO PICASSO	6	4	FRIDA KAHLO	7	4	PABLO PICASSO	0
5	VINCENT VAN GOGH	6	5	PAUL CÉZANNE	9	5	VINCENT VAN GOGH	0
6	CLAUDE MONET	1	6	HENRI MATISSE	6	6	CLAUDE MONET	0
7	FRIDA KAHLO	-3	7	CLAUDE MONET	-1	7	EL GRECO	1
8	GUSTAV KLIMT	1	8	EL GRECO	-1	8	MARC CHAGALL	4
9	EL GRECO	-1	9	GUSTAV KLIMT	6	9	DIEGO RIVERA	-2
10	JOAN MIRO	4	10	PABLO PICASSO	-6	10	PAUL KLEE	1
11	PAUL GAUGUIN	4	11	VINCENT VAN GOGH	-6	11	FRIDA KAHLO	-2

Table 19: First positions in one word⁴⁵

2008			2005			2001		
1	WASHINGTON	0	1	WASHINGTON	0	1	WASHINGTON	0
2	CLINTON	2	2	KENNEDY	5	2	CHRIST	1
3	DALÍ	34	3	LINCOLN	1	3	CLINTON	1
4	DISNEY	1	4	CLINTON	-1	4	LINCOLN	-2
5	LINCOLN	-2	5	DISNEY	0	5	DISNEY	0
6	CHRIST	3	6	JEFFERSON	2	6	NEWTON	0
7	KENNEDY	-5	7	NEWTON	-1	7	KENNEDY	2
8	MADONNA	27	8	EINSTEIN	5	8	JEFFERSON	-1
9	JEFFERSON	-3	9	CHRIST	-7	9	GORE	7
10	BACH	18	10	DARWIN	18	10	DALÍ	39

What has been learnt from these measurements?

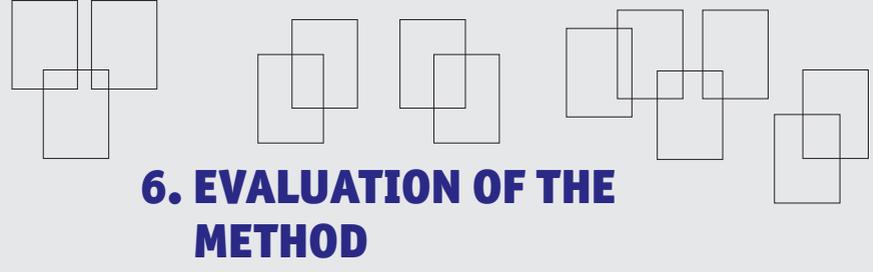
First of all, that whereas culture and business are closely associated as in music or cinema, the online bias towards American culture is obvious. However, in the themes where culture stands alone like in literature, science or graphic arts, the cultural representation on the Internet, as measured across personalities, is not biased. The

⁴⁵ Note that “bush” is not part of the sample. If it were, in 2008 it would have come second above “clinton”.

presence of French authors in literature or French researchers is as clear as the presence of Spanish-speaking painters. In the first two measurements, an initial handicap was sensed for French-related cultures, and much more for Spanish-related cultures. This was overcome in 2005. Since then, there has been no global change and that is why the 2008 measurement will be probably the last one to use that methodology.

Secondly, the Internet is a highly responsive medium that rapidly reflects real life events yet can equally 'lose interest' in particular people and events. This explains the surge and decline of some personalities whose online presences are boosted by an event (like the release of a movie on Che Guevara in 2008, sparking heightened interest in this personality) or more subtly, by some sociological trend making personalities more or less fashionable (it is interesting for instance to see the evolution over time of the respective cyber-fame of Albert Camus and Jean-Paul Sartre).

Thirdly, a kind of 'globalized culture' can be sensed on the Internet. It is very probable that this culture excludes important elements or personalities which are extremely relevant to local cultures but have not 'made it globally' or been deemed to have wider, global-reaching importance. Yet, returning to a consideration of language in the portrayal of cultural diversity, the issue of how minority cultures are represented online, is wide open. This includes cultures like those of indigenous people, that may not actually constitute a minority, yet still fall victim to the digital divide in terms of access to and uptake of ICTs.



6. EVALUATION OF THE METHOD

6.1 ITS UNIQUENESS AND ADVANTAGES

Objective consideration of the methodology and results described above reveals the following strengths:

- The method described above makes logical and productive use of extremely versatile online tools, namely search engines. Over the course of the stipulated research period, the only limitation in measuring word-concept citations in cyberspace was by way of their normatively bound connectivity to search engine indexes. Any question raised about the effectiveness of relying on such indexes to extrapolate across cyberspace, could be responded to with another question: what practical relevance does a Web page have if it is not indexed?
- Further, the research project has consistently and over an extended period of time been one of the very few amongst its peers to have transparently revealed its process, results and detailed method.
- Considering the valid argument that a perfectly neutral selection of appropriate word-concepts is impossible as far as culture is concerned, all precautions have been taken, both in terms of language and culture, to minimize biases. The practical list of word-concepts (and personalities) was chosen to maximise the credibility of the results.
- A formal, standardized statistical method was employed to achieve consistency for the series of 13 measurements undertaken from 1996 to 2008. This helped build additional project credibility.
- Compared to other methods used in other research projects measuring linguistic diversity online, this study has included the measurement of other spaces in addition to the Web. Further, the method enabled the project team to obtain more detailed results per language and per country, providing the only existing approximation of its kind, and leading to the creation of powerful indicators.

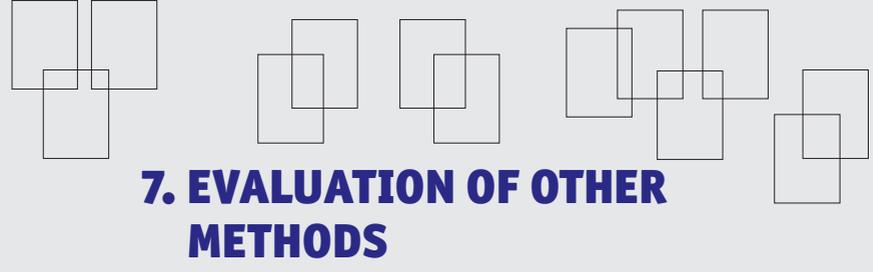
- So far, this study has been the only one to offer a consistent series of measurements, so as to give a broad perspective on the subject since all other research projects undertaken, as listed or discussed, were either once-off studies or short-term.

6.2 ITS WEAKNESSES AND LIMITATIONS

The methodology used and explained above does, however, have some weaknesses:

- It is limited to a small sample of languages, and the marginal cost to add a new language is relatively high. It is not a practical approach for the generalization of the weighting of most languages in the Web: algorithms of language recognition applied to data bases obtained by direct crawling⁴⁶ of the Web (as in the LOP) is, without a doubt, the standard method to be applied in future.
- It does not directly provide an absolute value for any language. The estimation of the absolute presence of English, which is used to compute other languages values, is made by a non-systematic process. This appears to be increasingly problematic due to changing search engine technology and the growing diversification of languages in cyberspace.
- It only measures the presence of a language in the indexed part of the search engines. This inconvenience was unimportant till 2005. Until then, as search engines covered more than 60% of the visible Web, extrapolation of measurement results was sensible. However, after 2005 this became a more serious impediment. Nowadays, the evolution of search engines necessitates a new method based on direct crawling and counting.
- Some of the indicators use figures which are either controversial (such as the number of speakers per language in the world) or quite unreliable (such as the number of speakers per language which are Internet users). This obviously impacts the confidence interval of some of the produced indicators.

⁴⁶ Crawling is the process of automatic and systematic browsing the Web pages (and possibly storing a representation of them) as it is done by search engines. See http://en.wikipedia.org/wiki/Web_crawling.



7. EVALUATION OF OTHER METHODS

A number of alternative methods for measuring the extent of languages present in cyberspace were used and published during the project's lifespan. Numerous publications of results were made by marketers, without a clear description of the methods used. What follows is a selection of what are considered the most relevant actions to the topic.

BABEL TEAM: A joint initiative from Alis Technologies and the Internet Society

This effort, which was presented as the very first although it was actually made several months after FUNREDES first studies, is from Alis Technology, a Canadian company, and was published with the support of Internet Society in June 1997. Although the report⁴⁷ announced two measurements would be made per year, it was a once-off activity. In spite of this, it is very interesting to analyze, as it offered the first trial of what would become the method used later by OCLC - twice in 1999 and 2002 - to support the media discussion about the steady presence of English on the Web at around 80% in the studied time-frame (see [How "World Wide" is the Web?](#) and [Trends in the Evolution of the Public Web: 1998 – 2002](#) in references).

The Alis method is based on a random sample of 8,000 Web sites on the home pages of which⁴⁸ an algorithm of language recognition is applied with the capacity to identify 17 different languages to obtain the language repartition and extrapolate to the whole Web. Before this occurs, a visual check is made for a subset of the sites, to locate the error rate of recognition. From there, some corrections can be applied to the results which are not discussed in the documentation.

The method is transparently explained and some of the data is even made publicly available such as the list of Internet Protocol

47 <http://alis.isoc.org/palmarens.en.html>

48 Actually the number of sites which were analyzed is a little above 3000.

(IP) numbers which were analyzed. The main inconveniences of the method has been first that it has not lived up to its promise of replication, and second that the results are published after a unique measurement. The second fact undermines any trust in the results⁴⁹. Its limitations are important to be clarified in relation to any evaluation of the OCLC project.

- 1) The method presupposes that the language employed on the Home Pages of websites fairly represent the language distribution for the whole Web. This is in spite of the fact that many sites in other languages have their home page in English regardless of the language used throughout the site, or are bilingual.
- 2) The algorithms of language recognition were not - and are still not - totally reliable although they have improved since 1997 and they tend to offered biased results in favour of English⁵⁰.

But those are minor limitations compare to the following two:

- 3) In terms of statistics, one can question the fact that the 3,000 servers selected randomly would accurately and proportionately represent the reality and diversity of a cyberspace universe of, at that time, around one million servers. In other words, how could a random sample of 0.3% of the total number of servers worldwide accurately reflect the diversity of the cyberspace universe? It is true that, for instance, surveys are able to forecast the results of an election process fairly precisely, but the sample is not made randomly. To the contrary, it is constructed to provide a fair representation of the topology of the whole elective population such as age, sex, location, socio-economic status, etc.

Furthermore, the sole, conventional manner to evaluate the working hypothesis has not been realized. This leads to an explanation of the next, last and largest limitation of this study:

- 4) In order to have a degree of statistical validity and scientific rigour, the process to be followed shall be a series of repeated measurements with a different random sample of IP numbers that enable

49 The results gave the percentage of English in the Web above 80%.

50 The Language Observatory Project reports an error rate of around 10%.

the analysis and distribution of the random variable, in order to understand its statistical behaviour as have been achieved in the FUNREDES/Union Latine method, by summing 57 concepts. However, this would have rendered manual verifications of the process too labour and time intensive.

OCLC Web Characterization Project

OCLC is a famous project in terms of measuring online linguistic diversity. It provides comprehensive services for librarians and has also provided consistent and reliable data on Internet demographics under the Web characterization project until 2003⁵¹. However, the data produced for the language presence on the Web copied and followed the Alis methodology, and suffered from the same limitations as those described above. It projected the same figure of 72% for the English presence on the Web in 1999 and 2002. Both measurements were made using the same methodology⁵².

The last publication made in 2002, concludes “that growth in the public Web, measured by the number of Web sites, has reached a plateau” and “there are no signs that this US-centric, English-dominated distribution of content is shifting toward a more globalized character” (see Trends in the Evolution of the Public Web: 1998 - 2002, in references).

At the same time, the FUNREDES/Union Latine study indicated English language has a presence of 50%, and there was an extremely visible trend of linguistic diversification of the Web. Being the unique, US-based source of information on the subject, and benefiting from the prestige of the OCLC name, this study supported the concept of English retaining a steady 80% presence. However, this was obtained using a flawed methodology and against all obvious and visible trends. Yet it was used as reference for media reports, conveying the story of a Web completely dominated by English.

Even over time, as it became evident that the demographics of the Web were rapidly evolving, and the English-speaking proportion of Internet users dropped from 60% to less than 30% between 1999

51 <http://www.oclc.org/research/projects/archive/wcp/default.htm>

52 <http://www.oclc.org/research/projects/archive/wcp/stats/intnl.htm>

and 2005. This statistic remains the main reference of papers as late as 2005, as in Paolillo's article in Measuring linguistic diversity on the Internet (see reference).

INKTOMI STUDY

In February 2000, Inktomi, one of the main search engine companies at the time⁵³, used very effective online marketing to circulate the results of its study⁵⁴ about the presence of English on the Web, with the following results:

Table 20: Results of Inktomi study

LANGUAGE	PROPORTION (%)
English	86.54
German	5.83
French	2.36
Italian	1.55
Spanish	1.23
Portuguese	0.75
Dutch	0.54
Finnish	0.50
Swedish	0.36
Japanese	0.34

The total of the percentages of languages mentioned in the table above reaches 100%, yet many other languages were actually present on the Web. This obviously impacts the real absolute value of English⁵⁵. As such, this marketing operation contributed heavily to the overstatement of English dominance of the Web in 2000. The media projection of this figure for English language prominence, at around 80%, was symptomatic of the general loss of scientific interest in the topic. An estimation of the proportion of web pages of other languages in addition to those mentioned above as being above 20% of the total, would effectively decrease the proportion of English to below 70%.

53 It was bought by Yahoo in 2002 and disappeared.

54 No reference on the used methodology was ever reported.

55 The percentages should have been computed in relation to the total number of languages or a proportion left in the table for "the others".

THE METHOD NAMED “COMPLEMENT OF THE EMPTY SPACE”

Starting in March 1998, with AltaVista and following later with Google, some search engines asked their users about their perceptions of the linguistic compositions of the search engine indexes. Asking to search an expression such as “-bhjvfvj” - meaning search for all except nothing and this is why it was named *method of the complement of the empty space*, e.g. the whole space -, the answer would be the whole index with the total of indexed pages. The same query for a given language would produce the estimation by the search engines about the number of pages in this language. Obviously, this method reflects the strong bias of language recognition algorithms towards English⁵⁶. It is merely to be taken as a first, gross approximation of the space occupied by different languages on the Web. In any case, this method was frequently used⁵⁷ to check the evolution of English and the growth of new languages on the Web. On July 3rd 2008, Google reflected this language repartition in its data base, as tabled below:

Table 21: Google estimation of web pages per language

LANGUAGE	TOTAL PAGES	PERCENTAGE
Arabic	340,000,000	0.68%
Bulgarian	169,000,000	0.34%
Catalan	46,400,000	0.09%
Chinese (simplified)	3,770,000,000	7.49%
Chinese (traditional)	796,000,000	1.58%
Croatian	113,000,000	0.22%
Czech	269,000,000	0.53%
Danish	249,000,000	0.49%
Dutch	583,000,000	1.16%
English	25,580,000,000	50.82%
Estonian	129,000,000	0.26%
Finnish	225,000,000	0.45%
French	1,750,000,000	3.48%
German	2,470,000,000	4.91%

56 To understand this bias it is sufficient to make few experiments of searching for a word in English and seeing the high percentage of pages of other languages which are mistaken for English.

57 Although, from time to time, the search engines stop offering this feature.

LANGUAGE	TOTAL PAGES	PERCENTAGE
Greek	148,000,000	0.29%
Hebrew	290,000,000	0.58%
Hungarian	278,000,000	0.55%
Icelandic	27,100,000	0.05%
Indonesian	132,000,000	0.26%
Italian	951,000,000	1.89%
Japanese	3,040,000,000	6.04%
Korean	968,000,000	1.92%
Latvian	43,200,000	0.09%
Lithuanian	95,600,000	0.19%
Norwegian	255,000,000	0.51%
Polish	675,000,000	1.34%
Portuguese	828,000,000	1.65%
Romanian	254,000,000	0.50%
Russian	1,470,000,000	2.92%
Serbian	61,800,000	0.12%
Slovakian	181,000,000	0.36%
Slovenian	97,500,000	0.19%
Spanish	2,180,000,000	4.33%
Sweden	116,000,000	0.23%
Turkish	835,000,000	1.66%
Armenian	2	0.00%
Byelorussian	959,000	0.00%
Esperanto	3,740,000	0.01%
Persian	116,000,000	0.23%
Tagalog	8,300,000	0.02%
Thai	418,000,000	0.83%
Ukrainian	69,100,000	0.14%
Vietnamese	301,000,000	0.60%
TOTAL	50,332,699,002	100.00%

A close follow-up of this information has allowed the team to verify that a large number of sources have used this method as a technique to project their original language estimation on the Web⁵⁸, without documenting the method. This started in 2000. In any case, since

⁵⁸ This is the case for example of VilaWeb, which has effectively marketed its results in 2001.

2005 the results presented by Google using that method are totally unreliable, as they vary considerably from one period of measurement to the next.

XEROX STUDY (2001)

A study was published in 2001 using an original linguistic technique, based on the frequency of commonly occurring words in a given linguistic corpus. This was meant to be able to predict the presence of particular languages on the Web (see [Estimation of English and non-English Language Use on the WWW](#) in references). The study presents results for 1996, 1999 and 2000 and therefore should be considered as the first historical study of language presence on the Web. The results produced in this Xerox study about language presence are all comparatively quite a bit lower than the percentages found in FUNREDES' study, as indicated in the following table:

Table 22: XEROX study results

RATIO TO ENGLISH	XEROX 10/96	XEROX 8/99	XEROX 2/2000	FUNREDES 8/2000
German	3.8%	7.1%	6.9%	11%
French	3.7%	5.4%	5.7%	7.33%
Spanish	1.7%	4.0%	3.9%	8.41%
Italian	2.0%	2.9%	2.8%	4.60%
Portuguese	1.7%	2.1%	2.4%	3.95%

THE LANGUAGE OBSERVATORY PROJECT (LOP)

Since 2003, the LOP has been run at Nagaoka University of Technology in Japan. This project has been developing a language identification machine which can identify language, script and encoding of a Web page using a statistical method common for text processing. It is said to have an error rate of 9%, with a scope of 350 languages that can be analysed. Using a search engine browser developed by the University of Milan, the LOP has collected and identified about 100 million Web pages under Asian (except China, Japan and Korea to avoid mass processing) and African ccTLDs in 2006 and 2007. This research shows that, in 2006 and 2007 respectively 40% and 41%

of Web pages were written in English in Asian ccTLDs, and 73% and 82% of pages in African ccTLDs (see [A Language and Character Set Determination Method Based on N-gram Statistics and Analysis of the Asian Languages on the Web Based on N-gram Language Identification](#), in references). In future, the project will be extended to ccTLDs of other regions or gTLD domains. The LOP has obtained the following results for the languages that are also the focus of the FUNREDES/Union Latine study as at 2007:

Table 23: The LOP study results

	English	Spanish	French	Italian	Portuguese	Romanian	German	Catalan
Asia	41.5%	0.02%	0.25%	0.01%	0.03%	0.03%	0.21%	0.04%
Africa	82.1%	0.11%	7.0%	0.07%	1.2%	0.03%	0.82%	0.04%

The LOP, a large consortium of scientists and experts of universities and organizations spanning 20 countries, certainly has the potential to evolve into a reference project for the measurement of language diversity on the Internet. Its current situation makes it a unique tool for the control of the presence of minority languages. If and when the Web browsing is applied to the whole Web universe, it would produce the result which everyone has been waiting for within the limitation of the language recognition algorithms, which are still close to an error rate of 10%.

UPC/IDESCAT

This project was started in 2003 by the Statistics Institute of the Catalan Government (IDESCAT) and realized by Universitat Politècnica de Catalunya (UPC). It collected a data base of some 30 million domain names, from which it extracted a subset of two million domain names. Language recognition algorithms were applied to these, as in the LOP project.

The project shows the following results for 2005, tabled below. These are relatively close to the results of the FUNREDES/Union Latine study.

Table 24: UPC results for 2005

LANGUAGE	UPC	FUNREDES
English	42.9 %	45 %
Catalan	0.16 %	0.14% (2007)
Spanish	3.4 %	4.60 %
German	6.2 %	6.94 %
French	6.2 %	4.95 %
Italian	7.9 %	3.05 %

However, the results for 2006 are quite different from those of the FUNREDES/Union Latine study:

Table 25: UPC results for 2006

LANGUAGE	UPC
English	71.8 %
Catalan	0.47 %
Spanish	2.14 %
German	14.5 %
French	3.8 %
Italian	0.7 %

CULTURES ON THE INTERNET

This research team has found only one other attempt to estimate cultural biases on the Internet. A literature review identified a group of Spanish researchers who researched this topic in 2003 (see *Iconos culturales hispanos en Internet* in references). The fact is, this work has just reused FUNREDES' simplistic methodology and tried to enhance it. The marginal improvement⁵⁹ does not seem worth the marginal cost increment⁶⁰ and the result does not shed any new light on the subject.

59 As mentioned, the method is too simple to reflect the complexity of culture. The few improvements made by the group of researchers did not manage to change that reality or provide a paradigm shift for the results.

60 The FUNREDES' research team was surprised after reading the paper and discovering that this group preferred to pay a private company to redo the programming FUNREDES had already undertaken, instead of opening communication and seeking collaboration with FUNREDES!

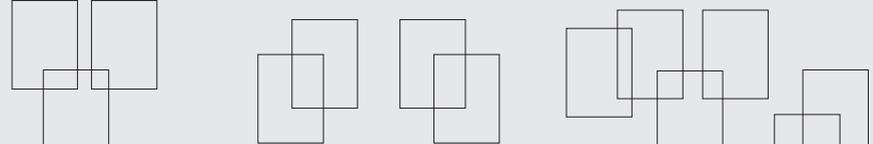
8. PERSPECTIVES

The field of linguistic diversity on the Internet is coming out of a period of difficulties and a lack of interest on behalf of international organizations and the academic world. The need for linguistic policies to defend or promote languages on the Internet is receiving more public awareness every day, and the need for reliable indicators naturally follows. Projects such as the LOP are indicating new approaches capable of producing a broad range of information on the presence of languages on the Internet. However the full requirements are much more complex and go beyond the statically represented proportion of languages in different cyberspaces. The usage of languages in email, chat and on websites is still an unknown element. However it is of continued importance, as it demonstrates the language dynamics in users' behaviour in the Internet⁶¹.

In that context, the FUNREDES/Union Latine study could be symptomatic of the 'prehistoric period' of measuring linguistic diversity on the Internet. The current evolution of search engines shows it has reached its limits in its actual form. Yet, the niche opportunity for an alternative method to cross-check the results of other methods which are relying on Language Recognition Algorithms still exists. In order to elicit meaningful results, it would, however, necessitate a redesign of the part of the methodology which is based on search engines.

In other words, such an evolution of the research method calls for a system of browsing the Web and undertaking the linguistic assessment directly during the browsing process. Counters should check the results at the end of the process. This would enable some kind of quality evaluation to be added, together with a word count. This heralds a promising field of research that has not yet been undertaken.

61 The idea of using approaches such as the one used by Alexa.com to weight the linguistic behavior of users would be a very promising one. Alexa installs a voluntary spyware in the personal computers of those people who agree, and reports on their navigation choices. From there, Alexa is able to compute interesting data from user behaviors, including a hit parade of visited web sites.

A decorative graphic at the top of the page consists of several overlapping squares of varying sizes and positions, creating a complex geometric pattern. The squares are white with black outlines and are set against a light gray background.

The feasibility and cost of this new, evolved method of research is currently under examination. FUNREDES is seeking a partnership, especially with reference to the web-crawling aspect of this potential project. Furthermore, it is intended, in collaboration with Antilles-Guyane University, to add French Creoles to the list of languages to be calculated and processed.

Regardless of the future of the FUNREDES/Union Latine method, the linguistic diversity of cyberspace is becoming an increasingly priority issue for building inclusive knowledge societies. This fact alone will create more need for professionally built indicators capable to help monitor language policies in cyberspace.

REFERENCES

Crystal, D. Language and the Internet. Cambridge University Press, 2001, ISBN-10: 0521802121, ISBN-13: 978-0521802123

Bergman, M.K. White Paper: The Deep Web: Surfacing Hidden Value. in Ann Arbor, MI: Scholarly Publishing Office, University of Michigan, University Library vol. 7, no. 1, August, 2001

<http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104>

Paolillo J.; Pimienta D.; Prado, D.; and ale. Measuring linguistic diversity on the Internet. UNESCO Institute for Statistics Montreal, Canada - UNESCO, 2005 (CI.2005/WS/06)

http://portal.unesco.org/ci/en/ev.php-URL_ID=20882&URL_DO=DO_TOPIC&URL_SECTION=201.html

Lavoie, B.F.; O'Neill, E.T. How "World Wide" is the Web? Annual Review of OCLC Research, 1999. <http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003496>

O'Neill, E. T.; Lavoie B.F.; Bennett, R. Trends in the Evolution of the Public Web: 1998 - 2002. D-Lib Magazine, 9.4., 2003

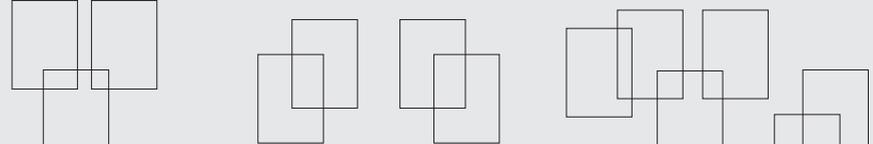
<http://www.dlib.org/dlib/april03/lavoie/04lavoie.html>

Grefenstette, G.; Noche, J. Estimation of English and non-English Language use on the WWW, Technical Report from Xerox Research Centre Europe, 2000 <http://arxiv.org/ftp/cs/papers/0006/0006032.pdf>

Suzuki I.; Mikami Y.; and ale. A Language and Character Set Determination Method Based on N-gram Statistics. In ACM Transactions on Asian Language Information Processing, Vol.1, No.3, September 2002, pp.270-279

Nandasara S.T.; and ale. Analysis of the Asian Languages on the Web Based on N-gram Language Identification. In The International Journal on Advances in ICT for Emerging Regions (ICTer), Volume 1, Issue 1, 2008

Monràs F.; and ale. Estadística de la presència del català a la xarxa d'Internet i de les característiques dels Webs Catalans, in Llengua i ús: Revista tècnica de política lingüística, ISSN 1134-7724, N°. 37, 2006, pags. 62-66



Cueto L.; Soler C.; Noya J. Iconos culturales hispanos en Internet (lo que ven los buscadores). in El Español en el Mundo: anuario del Instituto Cervantes 2004 / coord. por Paz Lorenzo, ISBN 84-01-37892-3 , pag. 127-190, 2004

<http://www.realinstitutoelcano.org/publicaciones/109/109.pdf>

Diki-Kidiri, M. Comment assurer la présence d'une langue dans le cyberspace? Unesco Cl.2007/WS/1. 2007

<http://unesdoc.unesco.org/images/0014/001497/149786F.pdf>

Pimienta D. Accessing content, in Global Information Society Watch, 2008, APC, ITEM, HIVOS Editors. <http://www.giswatch.org/gisw2008/thematic/AccessingContent.html>

(Tables footnotes)

- 1 The estimate of the population of speakers (first or second language) for the main spoken languages is:

X = NUMBER OF SPEAKERS (MILLION)	LANGUAGES
X > 500	CHINESE(S), ENGLISH, INDU(S)
200 < X < 500	SPANISH, RUSSIAN, ARABIC(S)
100 < X < 200	BENGALI, PORTUGESE, JAPANESE, INDONESIAN, GERMAN, FRENCH

- 2 <http://funredes.org/lc2005/L6/english/evol.html>
- 3 Source: Union Latine (2000)
- 4 This is the estimate of the world population. Note however, that the total number of speakers would be a higher figure, taking into account the number of people that speak more than one language.
- 5 30% would be a "guesstimate" of the population of the World that speaks more than one language. This figure is probably close to reality in OECD countries, but not so in many developing countries, where the average person used to speak 2 or 3 languages (like in Africa).
- 6 Source Internet Word Stats (2005)
- 7 This is the estimate of total Internet users.
- 8 This is the percentage of the world population that has Internet access.
- 9 Source: FUNREDES/Union Latine (2005)
- 10 Measured as the ratio of % Web pages per language by % Internet users per language.
- 11 Sum of the studied language except English.
- 12 Sum of the rest of the World languages.
- 13 In million. Source: Global Reach until 2005 and then Internetworldstats
- 14 A decrease of productivity with a similar content production percentage indicates a growth of Internet users not followed by a subsequent growth of content.

Secretariat
UNESCO
Communication and Information Sector
Information Society Division
1, rue Miollis
75732 Paris cedex 15
France

Tel.: + 33.1.45.68.45.00

Fax: + 33.1.45.68.55.83

www.unesco.org/webworld

Paris: UNESCO, 2010