

**Linguistic diversity on the Internet:
a critique of estimation techniques,
and published trends,
in online language usage**

Peter Gerrand

School of Historical & European Studies



pgerrand@gmail.com

**Presentation to Communications Policy & Research Forum
Sydney, 26 September 2006**

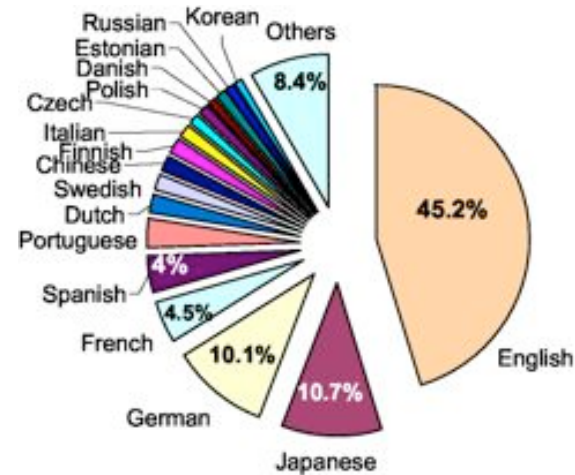
Public policy considerations in estimating online language diversity

- Funding Government online services
- WSIS - equity in the Internet
- The long-term survival of 'minority languages'
 - “Close to half of the 6,000 languages spoken in the world are doomed or likely to disappear in the foreseeable future” UNESCO 2002

OECD data: 2002 to 2005

OECD Figures for 'secure servers' in major languages:

Date:	Aug-02	Feb-05
English	57.6%	45.2%
Japanese	9.9%	10.7%
German	10.5%	10.1%
French	4.1%	4.5%
Spanish	5.3%	4.1%

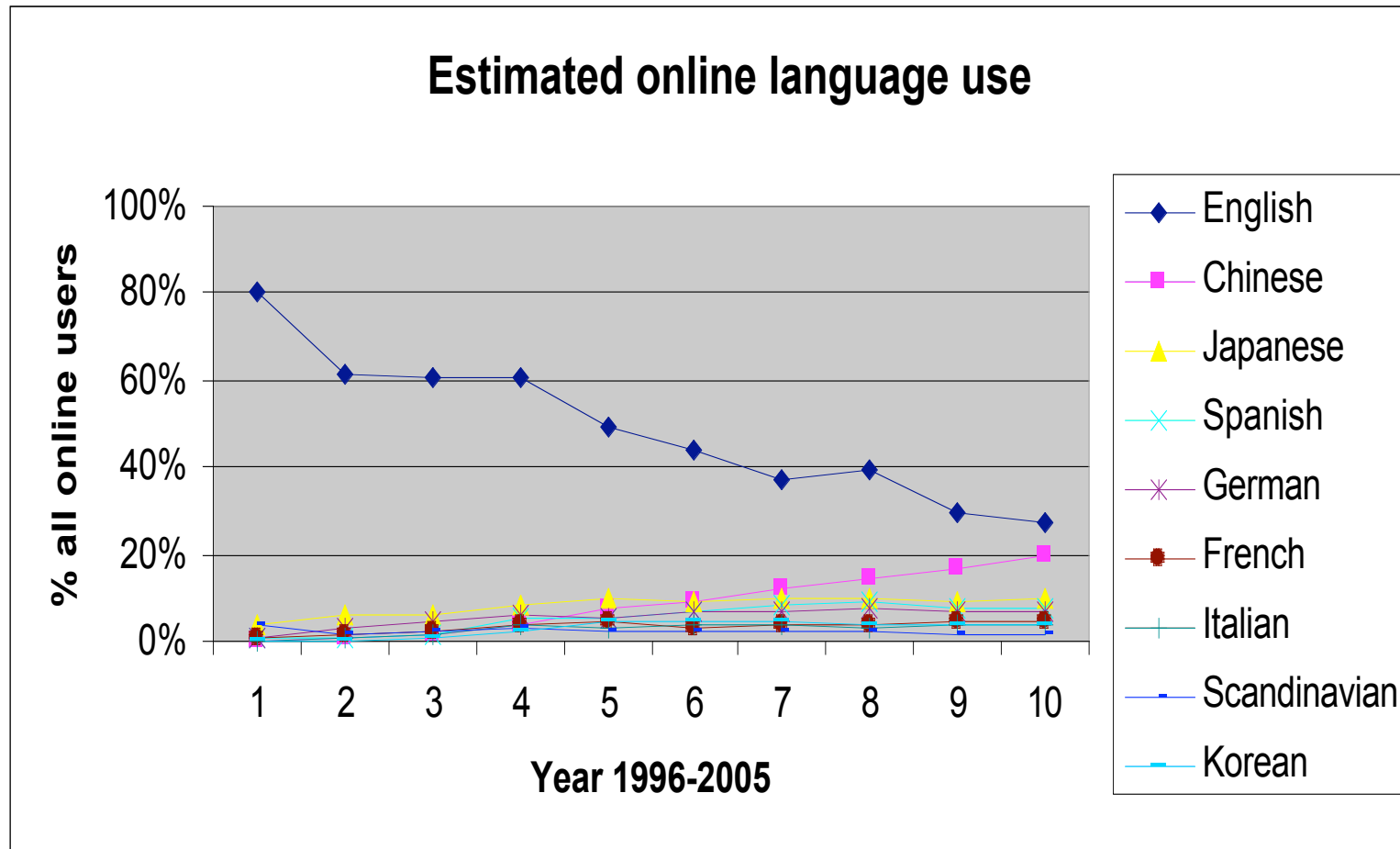


Feb-05 data

This is a measure of commercial use of languages on the Web

Global Reach's "Internet statistics by language"

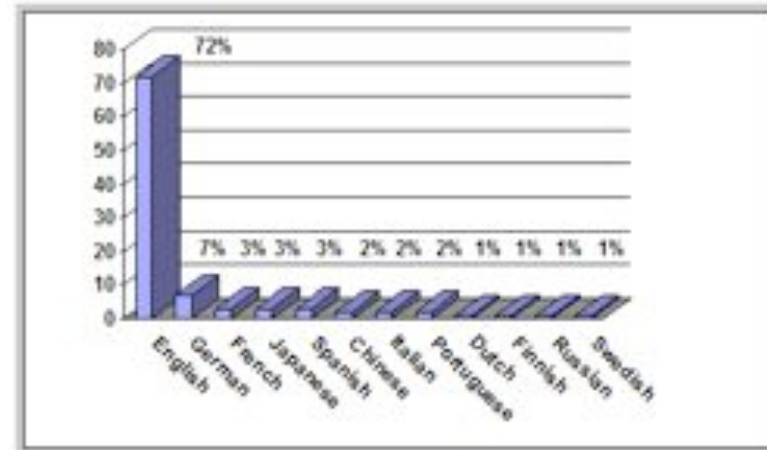
ref: www.global-reach.biz



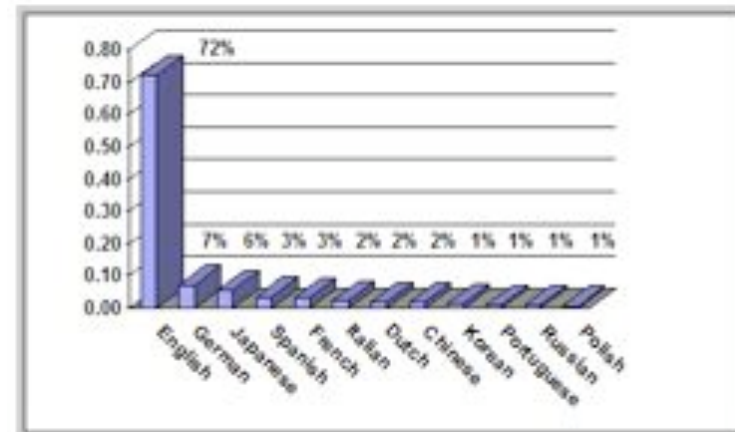
The paradox: relative position of English stays constant at between 70 and 80%

Studies by O'Neill et al (2003) at OCLC* in 1999 and 2002, based on direct measurement, show the relative frequency of English on the Internet as staying constant at 72%, well ahead of German, French, Japanese and Spanish

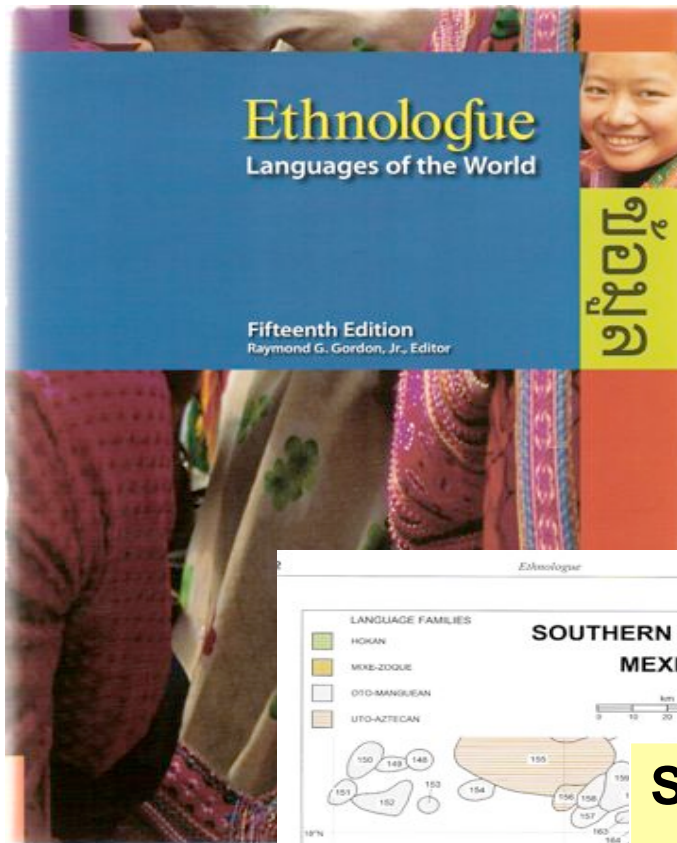
*Online Computer Library Centre, USA



1999 data



2002 data



Global Reach's estimates are based on Ethnologue data

6,912 living languages (2005) listed by country, names and by *ISO 639-2* 3-letter codes - and mapped for demography



Spanish (Español, Castellano, Castilian) [spa]
28,173,600 in Spain (1986).

Population total all countries: 322,299,171.

Central and southern Spain and the Canary Islands.

Also spoken in Andorra, Argentina, Aruba,

Australia, Belgium, Belize, Bolivia, Canada, [...]

Deficiencies of Ethnologue

Obsolete and asynchronous census data

- e.g. Ethnologue (2005) on English:

[UK:] **English** [eng] 55,000,000 (1984) *Lg Use:* National language. *Dialects:* [34 listed]

[USA:] **English** [eng] 210,000,000 (1984) *Dialect:* Black English. *Other:* many regional dialects

[Australia:] **English** [eng] 15,682,000 (1987). *Dialect:* Australian Standard English, Aboriginal English, Neonyungar. *Lg Use:* National language. *Other:* minor regional dialect differences

[India:] **English** [eng] *Lg Use:* National language. 2nd-language speakers: 11,021,610 (1961 census)

Ethnologue (2005) data on Spanish:

[Spain:] **Spanish** (Español, Castellano, Castilian) [spa] 28,173,600 (1986).

[Mexico:] **Spanish** (Español, Castellano)
[spa] 86,211,000 (1995) ...

[Argentina:] **Spanish** [spa] 33,000,000
(1995) Official language. See main entry
under Spain.

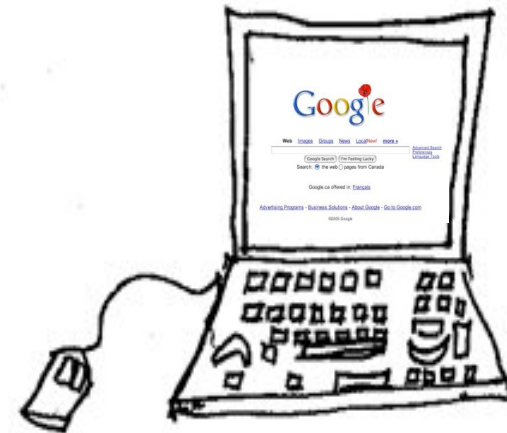
Proposal: a simple taxonomy for estimation techniques for online linguistic diversity

User activity



Web presence

User profile



Measuring user activity



- Major online activities are: email, instant messaging & chat groups; surfing the Web incl. blogs; Voice over IP; streamed audio or video; computer games
- **Direct measurement** of language usage over *all* Internet activity is beyond the resources of any – except the national security agencies – who have yet to publish their results.
- The only published studies of **direct measurement** of language usage have been studies of diglossia by users of email in specialised university email groups
 - e.g. Climent 2004 (on Catalan vs Spanish by online students at UOC, Barcelona); Durham 2004 (on German vs English by online Swiss medical students)
 - Simply not generalizable to larger populations of users

Global Reach's algorithm can be deduced as

$$i_{xy} = \sum_z s_{xyz} \left\{ \frac{a_{yz}}{p_{yz}} \right\} \quad \text{where in year } y$$

i_{xy} = **internauts speaking language x**

s_{xyz} = **speakers of language x in country z**

a_{xyz} = **Internet accesses in country z**

p_{yz} = **population in country z**

i.e. Total Internauts using language x

$$= \sum_z \{ (\text{speakers of } x)_z * (\text{per capita Internet access})_z \}$$

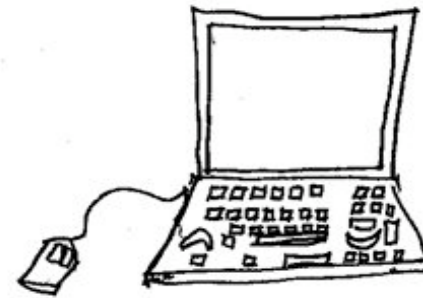
Problems with the Global Reach methodology:

- 1) In reality, Internet access varies with the language (e.g. aboriginal languages vs English in Australia) **therefore '3rd world' languages become over-estimated on the Net**
- 2) It is based on Ethnologue language population data that are often > 10 years out of date;

More problems with the methodology: It

- 3) includes some 2nd language data (bilingualism) for English, but apparently in the US alone
- 4) ignores the use of English, Spanish etc as 3rd or 4th languages
- 5) tends to under-estimate the online use of English, Spanish and Russian, and would over-estimate the online use of all minority languages – except that
- 6) the only minority language quoted is Catalan!

Measuring Web Presence



- (1) **Directly:** by automated language analysis of text found on a randomly addressed sample of all websites {Babel 1997; Lavoie 1999; O'Neill 2003}
- **Weakness:** only 29 languages detectable by automated technique used
- **Results:** English stayed constant at 72% from 1999 to 2002

(2) Using search engines

Deficiencies of search engines

- “Overlap studies show that about half of the pages in any search engine database exist only in that database” (UC Berkeley 2006: The best search engines)
- “Search engines typically employ a variety of **proprietary indexing methods that are not open to inspection**, and these may bias the page counts returned in ways that cannot be corrected or even reckoned. A word need not appear in the page at all for it to be included in the count,” (Paolillo 2005)
- **Language restrictions:** Up to 2003, *AllTheWeb* could detect and index 48 different languages, including Catalan, Basque, Galician, Welsh, Faeroese and Friesian . It was then on-sold to Yahoo. Current mainstream search engines *Google*, *Yahoo*, *Teoma* can only distinguish between 36, 37 and 7 languages respectively – incl. only one minority language: Catalan..

Best research on minority language web presence

- In March 2002 by Xavier Gómez Guinovart, Uni. Vigo {Guinovart 2003}, and 18 months later by Jordi Mas i Hernàndez, Softcatala {Mas 2003}, both using *AllTheWeb*.
- Despite *AllTheWeb*'s database having grown 3-fold between their tests, their results are highly consistent. The 'top 8' listed languages are the same, with English leading at 60%, and Spanish rising from 8th (2.3%) to 5th (3.1%).
- In both lists, Catalan scored 23rd (0.09% to 0.14%); Galician (37th) and Basque (38th) dropped to 39th and 40th over the 18 months.

UNESCO initiatives

- World Summits on the Information Society
 - Geneva December 2003; Tunis November 2005
- Tunis Commitment includes:
 - “We further commit ourselves to promote the inclusion of all peoples in the Information Society through the development and use of local and/or indigenous languages in ICTs.”
- Atlas of the world's languages in danger of disappearing (2001)
- B@bel project (2002+)
- Multimedia resources for multi-language learning
- “Measuring linguistic diversity on the Internet” (2005): report by Paolillo et al, commissioned for Tunis'05



Conclusions

- The paradox between contradictory estimates of the relative use of English and other living languages on the Internet is resolved
- A simple taxonomy of methodologies between those that measure **user activity**, **user profile** or **web presence** shows that they are measuring quite different indicators.
- All have weaknesses, exposed in my PhD thesis.
- In particular, the use of Ethnologue data for user profile estimates is fraught.
- And the mainstream search engines are no longer convenient for measuring web presence of minority languages.

Recommendation

- Given UNESCO's Tunis Commitment, it would be appropriate for UNESCO to fund and co-ordinate the development of accurate estimates of language use on the Internet:
 - *not* by **user activity** (privacy issues) but by
 - **web presence** (by direct sampling), and by
 - **user profile** (the potential users)



Acknowledgment:

Drawings by Elizabeth Darling



Appendix: Results by Guinovart (2002) and Mas (2003), using AllTheWeb

Order	Language	Web Pages (M)	% total
1	English	442	60.73
2	German	51.2	7.035
3	Japanese	43.2	5.926
4	Chinese	26.2	3.600
5	French	24.6	3.379
6	Korean	20.4	2.799
7	Russian	19.6	2.689
8	Spanish	16.4	2.254
9	Italian	15.1	2.077
10	Portuguese	12.5	1.718

	Language	Web Pages (M)	% total
1	English	1,280	60.42
2	German	182.0	8.591
3	French	99.7	4.708
4	Japanese	69.7	3.291
5	Spanish	65.8	3.107
6	Chinese	65.7	3.103
7	Korean	64.6	3.050
8	Russian	42.3	1.996
9	Italian	41.8	1.975
10	Dutch	41.1	1.941