Measuring the Web in Portuguese

Rachel Aires   Diana Santos

Linguateca

SINTEF Tele og Data

Pb 124, Blindern
NO-0314 Oslo, Norway
Rachel.V.Aires@sintef.no, Diana.Santos@sintef.no

It has been stated that Portuguese is the sixth language in the world in terms of native speakers, the fourth most used in Internet Interaction, and that Portuguese native speakers constitute 3% of Internet population. Our focus is on the finer characterisation of this increasingly large community of users in terms of: sheer size, user population, content distribution, parallel content and search engine coverage.

Size in pages and words

A. Queries with several of the most frequent grammatical items in Portuguese

We tried 7 queries with sizes between 10 and 5 words and got 6 queries that roughly represent Portugues. We present the results for two of them:

1. The one that returned the minimum number of pages

   que AND o AND e AND do AND da AND em AND um AND para

2. The one that returned the maximum number of pages

   de AND a AND o AND do AND da AND e

	www.alltheweb.com	www.altavista.com	www.google.com
Query 1	10,790,800	2,537,767	2,790,000
Query 2	20,807,956	7,152,022	4,260,000

B. Common words

We used 3 queries with content words that are common in Portuguese.

Queries	www.alltheweb.com	www.altavista.com	www.google.com
homem	2,682,556	507,529	705,000
filho	1,990,393	519,099	577,000
povo	1,532,175	320,669	380,000

C. Low frequency words

We performed queries with 8 low frequency words: *austero*, *arrumação*, *cara-dura*,   *cara(s)-de-pau*,   *cegonha(s)*,   *guarda-no(c)turno*,   *gato-pingado*, *matrimônio / matrimónio*.

Query	www.alltheweb.com	www.altavista.com	www.google.com
Gato-pingado	145	67	169

The aim of the experiments B and C was to estimate the size of the web in Portuguese using the relative frequencies of the words in Portuguese corpora as estimators of the percentage of the Web covered. The final estimation remains to be done, however, since while the frequency of grammatical words increases linearly with text size, the picture is different for content words, which tend to appear in bursts. So words like *cegonha* or *arrumação* tend to appear in texts about these subjects very probably more than once. So, one has to correct their relative frequency by a factor that models this (for example 1/2.5), and this factor is probably lexically dependent as well (may vary with part of speech and the words themselves).

D. Words belonging both to English and Portuguese

We made 7 queries at www.alltheweb.com, using this engine's language facility: *Israel*, *Shakespeare*, *Timor, legal*, *Portugal*, *Jorge Amado*, *Eça de Queiroz*. It returned from 4,61 to 46,55 times more documents in English than in Portuguese, except for the last two where the picture was reversed.

Query	www.alltheweb.com - English	www.alltheweb.com - Portuguese	…
Timor	2,251,659	84,632	26,61

E. Corresponding distinct words in English and Portuguese

We then input to the same engine 17 pairs of queries: *perigoso*/dangerous*, *fidelidade*/fidelity**, *cavalo*/horse**, *universidade*/university**, and one for each month (*Janeiro/January*, …). The search engine returned from 3.18 to 48.85 times more documents in English than in Portuguese

Query in English	Query in Portuguese	…
fidelidade/fidelidades = 263,165	Fidelity/fidelities/ faithfulness/faithfulnesses = 837,476	3,18

F. Reproduction of Grefenstette´s estimation method

Replicating Grefenstette's estimation method and Portuguese words (*com*, *uma*, *os*, *não*, *ao*, *mas*, *muito*, *seu*, *são*, *eu*, *foi*, *você*, *ele*, *pela*, *quando*, *pode*, *brasil*, *seus*, *um*) on Altavista, we got **5,090,230,228** words in early November 2002.

Parallel content

Following Resnik, we have looked for pages containing at least one hyperlink where "English" appears in the text or URL associated with the link, and at least one such link for "Portuguese". We got **4,246,916** pages.

But with these queries we also found dictionaries, automatic translators, language courses and pointers to products (books) in other languages. So, we looked for pages in Portuguese containing at least one hyperlink where "expression A" appears in the text associated with the link (and the other way around for pages in English). This was we found **300,303** pages in Portuguese that have a parallel version in English and **105,365** pages in English that have a parallel version in Portuguese.

Expression A: "english version" or "version in english" or "this page in english" or "this homepage in english" or " versão em inglês" or "versão inglesa", "esta página em inglês".

Content distribution (subject, size and type)

We investigated the content distribution on 5 subject areas: technical texts, news, cooking, dating, and sales on the web. We got **131,226** technical pages and **69,431** pages with recipes. We don´t yet have a definitive for the three remaining subjects because for them we used many query expressions that are not necessarily independent. An interesting research question is experimentally assigning weights to attribute to the different expressions we found out to be relevant to identify genre or content.
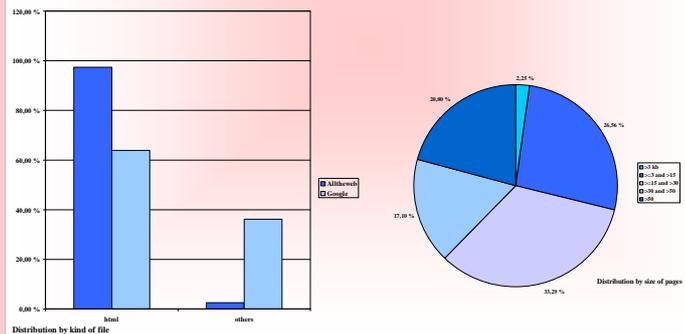
Search engine coverage and evolution

Our experiments showed that Alltheweb has the biggest database of indexed pages in Portuguese. Google comes second, followed by Altavista.

Based on the query 2 and on the information that alltheweb searchs on 2,095,568,809 pages and Google on 3,000,000,000, the percentage of Portuguese content on the indexed pages of these search engines can be estimated as 0.99% and 0.14% respectively.

 Comparing the Portuguese size of Altavista in X with December 2002, we observe that it grew % .

Remaining work

These are just preliminary results, that require further processing and the confirmation and independent verification of the estimation clues. We intend to do more experiments, and on a regular basis.

We would also like to asnwer the question "Who accesses Web content in Portuguese?" Some hints on how this might partially be done are to count how many references there are in other domains to Web in Portuguese; if link statistics are made available by public multilingual sites, count how many times Portuguese pages are accessed instead of the corresponding in other languages.

Explore better the content distribution investigating documents in Portuguese about other subjects, for example health, and compare it to general (all languages) content distribution.

We also want to investigate the coverage, strenghts and weaknesses of Portuguese-dedicated search engines (todobr, tumba, etc.) compared to general ones.