Measuring the Web in Portuguese
Rachel Aires  Diana Santos
Linguatex
SINTEF Tele og Data
Pb 124, Blindern
NO-0314 Oslo, Norway
Rachel.V.Aires@sintef.no, Diana.Santos@sintef.no

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 Portuguese. 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
   de AND a AND o AND do AND da AND e
   Query 1  10,790,800  2,577,767  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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>homem</td>
<td>2,682,556</td>
<td>507,529</td>
<td>705,000</td>
</tr>
<tr>
<td>filho</td>
<td>1,990,393</td>
<td>519,099</td>
<td>577,000</td>
</tr>
<tr>
<td>povo</td>
<td>1,532,175</td>
<td>320,669</td>
<td>380,000</td>
</tr>
</tbody>
</table>

C. Low frequency words
We performed queries with 8 low frequency words: austeria, arrumação, cara-
dura, cara(s)-de-pau, cegonha(s), guarda-neto(turno, gato-pingado, matrimônio / matrimônio.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gato-pingado</td>
<td>145</td>
<td>67</td>
<td>169</td>
</tr>
</tbody>
</table>

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, Portuguese, Jorge Amado, Éca de Queirós. 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.

<table>
<thead>
<tr>
<th>Query</th>
<th><a href="http://www.alltheweb.com">www.alltheweb.com</a> - English</th>
<th><a href="http://www.alltheweb.com">www.alltheweb.com</a> - Portuguese</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timor</td>
<td>2,251,659</td>
<td>84,632</td>
<td>26,61</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Query in English</th>
<th><a href="http://www.alltheweb.com">www.alltheweb.com</a></th>
<th><a href="http://www.alltheweb.com">www.alltheweb.com</a> - Portuguese</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>fidelidade/fidelidades</td>
<td>263,165</td>
<td>Fidelity/fidelidades/faithfulness/fidelity/faithfulness ses = 817,476</td>
<td>3.18</td>
</tr>
</tbody>
</table>

F. Reproduction of Grefenstette’s estimation method
Replicating Grefenstette’s estimation method and Portuguese words (com, uma, os, não, aha, mas, muito, seu, são, eu, foi, você, ele, pela, quando, pode, brasile, 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 “expression A” 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 searches 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 answer 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 lin statistics are made available by public multilingual sites, count how many time 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 (al languages) content distribution.

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