EuroGOV: Engineering a Multilingual Web Corpus

Börkur Sigurbjörnsson[1], Jaap Kamps[1,2], and Maarten de Rijke[1]

[1] ISLA, Faculty of Science, University of Amsterdam
[2] Archives and Information Science, Faculty of Humanities, University of Amsterdam
{borkur,kamps,mdr}@science.uva.nl

Abstract. EuroGOV is a multilingual web corpus that was created to serve as the document collection for WebCLEF, the CLEF 2005 web retrieval task. EuroGOV is a collection of web pages crawled from the European Union portal, European Union member state governmental web sites, and Russian governmental web sites. The corpus contains over 3 million documents written in more than 20 different European languages. In this paper we provide a detailed description of the EuroGOV collection.

1 Introduction

The world wide web is a natural setting for cross-lingual information retrieval. This is particularly true in Europe: many European searches are essentially cross-lingual. For instance, when organizing to travel abroad for a business trip or a holiday, planning and booking usually involves digesting pages in foreign languages. Similarly, looking for information about European culture, education, sports, economy, or politics, usually requires making sense of web pages in several languages. A case in point is the current European Union, which has no less than 20 official languages.

The linguistic diversity of European content is "mirrored" by the fact that European searchers tend to be multilingual. Some Europeans are native speakers of multiple languages. Many Europeans have a broad knowledge of several foreign languages, while English functions as the lingua franca of the world wide web. Moreover, many Europeans have a passive understanding of even more languages.

In view of the linguistic diversity of the European web and its searchers, a cross-lingual web retrieval task, called WebCLEF, was launched at CLEF 2005 [3]. Cross-lingual web retrieval requires a new document collection to be constructed, containing web content in many languages. Of course, there are many options for creating such a collection. Multi-lingual documents are abundant on the web. We have chosen to focus on pages of European government-related sites, where collection building is less restricted by intellectual property rights. The resulting collection, which we think of as a European counterpart of the `.GOV` collection [2], is called EuroGOV and has been made available in

Table 1. List of top-level domains covered in the EuroGOV collection. (Left): Domains which which were considered more important, based on previous/current CLEF interests. (Right): Other domains contained in the collection.

Main domains		Additional domains	
Domain	Country	Domain	Country
.cz	Czech Republic	.at	Austria
.de	Germany	.be	Belgium
.es	Spain	.cy	Cyprus
.eu.int	European Union	.dk	Denmark
.fi	Finland	.ee	Estonia
.fr	France	.gr	Greece
.hu	Hungary	.ie	Ireland
.it	Italy	.lt	Lithuania
.nl	The Netherlands	.lu	Luxemburg
.pt	Portugal	.lv	Latvia
.ru	Russia	.mt	Malta
.se	Sweden	.pl	Poland
.uk	United Kingdom	.si	Slovenia
		.sk	Slovakia

January 2005 [5]. The crawled pages were cleaned-up and organized in a uniform format, bundled and compressed down to manageable sizes. The collection is available under an individual or organizational license restricting its usage to research only; see [5].

In this paper we describe the EuroGOV collection in detail. The paper is organized as follows. We start by describing the crawling process in Section 2. Section 3 then lists various characteristics of the resulting collection, including the domains, the languages, and the link structure. We conclude with some discussion and future outlook in Section 4.

2 Crawling

Our initial plan for building EuroGOV was to obtain a focused crawl from the European Union seed `.eu.int`, and branch into the individual member states' governmental sites. However, restricting a crawler to government-related sites proved highly non-trivial. There is no simple way to tell a European government site apart from any other European site. For some government sites the crawling is smooth and we can easily filter out governmental pages (notable examples include `.gov.uk` and `.regeringen.se`). Most governmental sites, however, have more complex structures, and we could only focus the crawl by providing an explicit list of domains. As an example, we initially crawled 13 different domains to gather pages from the Finnish government. As the following domain list shows, there is no easy way of identifying Finnish governmental domains:

defmin.fi, formin.finland.fi, intermin.fi, ktm.fi, minedu.fi, mmm.fi, mintc.fi, mol.fi, om.fi, stm.fi, vm.fi, vnk.fi, and ymparisto.fi

These differences in domain naming traditions make it difficult to guarantee completeness of the information crawled for some governments. As a result, what we should realistically aim for is that EuroGOV contains the fairly complete content of

- the main government portals, and
- the main ministries

of the countries whose information we want to include in the corpus.

Our crawling process can be divided into three parts. Our initial seed was made by picking 2–3 main governmental sites for each EU member state. The seed contained 40 URLs and was created by referring to a list from the EU portal.[3] After completing several cycles of this crawl we realized that due to the varying structure of governmental sites, the portion of governmental pages covered differed considerably from one country to another. In order to try to get a better harmony in coverage we began a new crawl, now starting with a seed consisting of a list of ministries for a subset of the EU countries. The subset covered 12 countries and was chosen according to CLEF interests. The left column of Table 1 shows the list of main domains. The second seed list consisted of 131 ministries from 9 EU member states (the UK, Sweden, and the EU itself were considered adequately covered in the initial crawl). The seed was created by browsing the main government portals. The third crawl was performed when interest was expressed in including Russian government pages in the crawl. The Russian crawl was created from a single seed: www.gov.ru. The final collection was created by combining the three crawls into a single collection of pages.

3 EuroGOV Collection Characteristics

In this section we provide various statistics concerning the collection, including the domains covered, the languages it contains, and its link structure.

3.1 Domains

The EuroGOV collection has pages from the 27 primary domains listed in Table 1. There is a set of 13 main domains, shown on the left-hand side of Table 1, chosen in accordance with current CLEF interests and plans. We have attempted to include a sufficiently large number of pages from these 13 main domain. There are 14 additional domains, shown on the right-hand side of Table 1, from which pages are also included in the collection. The coverage of these additional domains is often less complete than the coverage of the main domains. Note that pages in the languages of the additional domains will 'creep in' anyway. For

[3] URL: http://europa.eu.int/abc/governments/index_en.htm

Table 2. Statistics of the EuroGOV collection over primary domains.

Domain	Pages				Size
	Total	Duplicated	Duplicates	Unique	(compressed)
.at	10,065	457	950	9,115	24M
.be	69,011	819	2,066	66,945	115M
.cy	1,972	52	52	1,920	7.9M
.cz	324,496	10,808	25,915	298,581	519M
.de	444,794	1,682	4,658	440,136	1.1G
.dk	2,144	497	519	1,625	5.4M
.ee	16,768	486	3,960	12,808	44M
.es	35,168	3,372	9,297	25,871	298M
.eu.int	374,484	32,838	58,415	316,069	1.9G
.fi	661,559	5,815	85,289	576,270	1.3G
.fr	156,450	11,144	21,894	134,556	545M
.gr	303	10	15	288	416K
.hu	330,822	361	1,082	329,740	1.5G
.ie	12,754	1,431	1,982	10,772	32M
.it	89,836	10,056	17,011	72,825	324M
.lt	10,765	751	1,131	9,634	8.8M
.lu	8,521	52	837	7,684	33M
.lv	317,404	10,357	25,711	291,693	675M
.mt	13,991	1,300	1,372	12,619	57M
.nl	149,949	6,097	18,911	131,038	434M
.pl	66,885	3,746	4,889	61,996	330M
.pt	147,445	2,454	8,744	138,701	753M
.ru	104,659	10,676	20,049	84,610	479M
.se	102,457	2,506	15,068	87,389	155M
.si	12,434	73	224	12,210	27M
.sk	58,020	3,288	3,764	54,256	128M
.uk	66,345	1,688	2,987	63,358	331M
Total	3,589,501	122,816	336,792	3,252,709	11G

example, the eu.int domain has ample pages in all of the 20 official languages of the European Union.

The EuroGOV collection features more languages and countries than are being used in the WebCLEF 2005 evaluation tasks. We made a deliberate choice to go for this extended list of countries and domains. This will facilitate future task extensions for cross-lingual web retrieval, or re-use of the collection for other purposes. We also feel that this reflects the natural situation when building a 'European' search engine.

The EuroGOV collection contains a total of 3,589,501 pages, and can be compressed in 11 gigabytes of data. Table 2 gives the page counts for each of the primary domains in the collection. The first column lists the primary domains in the collection. The second through fifth columns list the total number of web pages per domain; the number of MD5 checksums (of the page's content)

Table 3. Breakdown of the EuroGOV collection over document languages.

EuroGOV Collection		Domain `.eu.int`.	
Language	Percentage	Language	Percentage
finnish	20.28%	english	33.26%
german	18.20%	french	18.08%
hungarian	12.58%	german	9.08%
english	10.16%	finnish	6.24%
latvian	8.80%	spanish	5.75%
french	6.98%	dutch	5.29%
swedish	5.32%	danish	5.13%
portuguese	3.93%	portuguese	4.47%
dutch	3.91%	swedish	3.26%
polish	2.14%	greek-iso8859-7	2.92%
italian	1.70%	italian	2.64%
spanish	1.39%	latvian	1.13%
czech-iso8859_2	1.13%	polish	1.05%
slovak-windows1250	0.89%	estonian	0.60%
russian-windows1251	0.60%	lithuanian	0.51%
danish	0.49%	hungarian	0.40%
estonian	0.39%	czech-iso8859_2	0.05%
russian-koi8_r	0.30%	slovak-windows1250	0.04%
slovak-ascii	0.27%	romanian	0.03%
greek-iso8859-7	0.27%	slovak-ascii	0.02%
lithuanian	0.19%	russian-koi8_r	0.02%
irish	0.03%	icelandic	0.01%
welsh	0.01%	russian-windows1251	0.01%

that occur more than once; the number of pages that have a repeated MD5 checksum (and thus the same content as another page); and the number of unique pages. The final, sixth, column lists the total size of the pages when compressed. The five domains with the largest numbers of pages are: Finland (661,599), Germany (444,794), European Union (374,484), Hungary (330,822), Czech Republic (324,496). Although the number of pages per domain varies between 661,559 (Finland) and 303 (Greece), the number of pages is generally sufficient to support the building of a test collection. Specifically, the smallest set of pages for one of the main domains is 35,168 (Spain). It is unclear, at this point, to what extent the varying numbers of pages per domain is a result of the available web content, different link structure of different governmental sites, or of our particular choices in crawler software or seed points.

3.2 EuroGOV Language Distributions

What is the distribution of languages in the collection? To answer this question, we applied the TextCat language identification tool [4], which is based on [1], using a restricted set of 30 language models covering the European languages

Table 4. Breakdown over document languages for selected domains in the EuroGOV collection .

Domain .be.		Domain .de.		Domain .fi.		Domain .fr.		Domain .uk.	
Lang.	Perc.	Lang.	Perc.	Lang.	Perc.	Lang.	Perc.	Lang.	Perc.
french	36.78%	german	97.70%	finnish	81.15%	french	94.25%	english	99.05%
dutch	24.32%	english	1.37%	swedish	11.52%	german	2.49%		
german	21.61%	french	0.74%	english	7.26%	english	2.24%		
english	16.74%					spanish	0.81%		

only. Table 3 shows the results for the whole EuroGOV collection, as well as a breakdown for the `.eu.int` domain. Since pages may have little text or mixed language content, language identification may show multiple languages. For over two-thirds of the pages, a single candidate language stands out sufficiently clearly. Below, we analyze the language distribution on these pages.

When looking at the distribution of languages over the whole collection, shown on the left-hand side of Table 3, we see that the most frequent languages are Finnish (20%), German (18%), Hungarian (13%), English (10%), and Latvian (9%). It is a surprizing outcome that languages of the Finno-Ugrian family dominate the collection! The distribution of languages over the collection closely corresponds with the number of pages per domain (in numbers of pages, Finnish ranked first and Hungarian ranked fourth, see Table 2).

A look at the distribution of languages for Germany, France, and the UK, shown in Table 4, confirms this strong correlation between country and official language: In the German domain, 98% of the pages is in German. In the French domain, 94% of the pages are in French, and in the UK domain, 99% of the pages are in English. In countries with more than one official language, such as Finland (with Finnish and Swedish) or Belgium (with Dutch, French, and German), we see more language diversity within the corresponding domains. The language distribution for the Finnish and Belgian domains is also shown in Table 4. Since the languages and domains seem to be closely tied together, the distribution of the mixed language domain of the European Union, shown in Figure 1 and on the right-hand side of Table 3, is of great interest. Here, we see that English is the most used language, accounting for 33% of the pages, followed by French (18%) and German (9%). In Appendix A, the language distribution of each of the main top-level domains is given.

3.3 Link Structure

Table 5 lists a number of salient features of the EuroGOV link structure. The second and third columns give the counts of the number of links and the number of realized links (ones whose targets are in EuroGOV; columns 4 and 5 provide the average number of links per page and the average number of realized links per page. The last row but one provides averages over all top level domains, and the last row provides the total number of links in the collection.

The largest numbers of links can be found in the domains with the largest numbers of pages; just under half of the links are realized in the collection. The average number of links per page varies considerably between domains, and the average realized number of links is just over half of the average number of links, although the relative gap between the two numbers varies quite a lot between domains, (e.g., for `it` the average realized fan-out is 84% of all links per page, while it is only 16% for `ru`).

4 Discussion

EuroGOV was thought of as an experimental collection for evaluating cross-lingual web retrieval. As such, the collection serves its purpose well. However, EuroGOV has several limitations which should be taken into account when working with the current collection and planning for possible future extensions of the collection.

- *Completeness*: Quite some effort was put into collecting lists of governmental sites to crawl. This is, however, not a complete list. Especially for the Spanish and Portuguese domains, the collection contains only a very small fraction of the available pages on government-related web sites.
- *Incomplete description of the link structure*: A full link analysis of the EuroGOV collection has not been performed yet. This is not an inherent

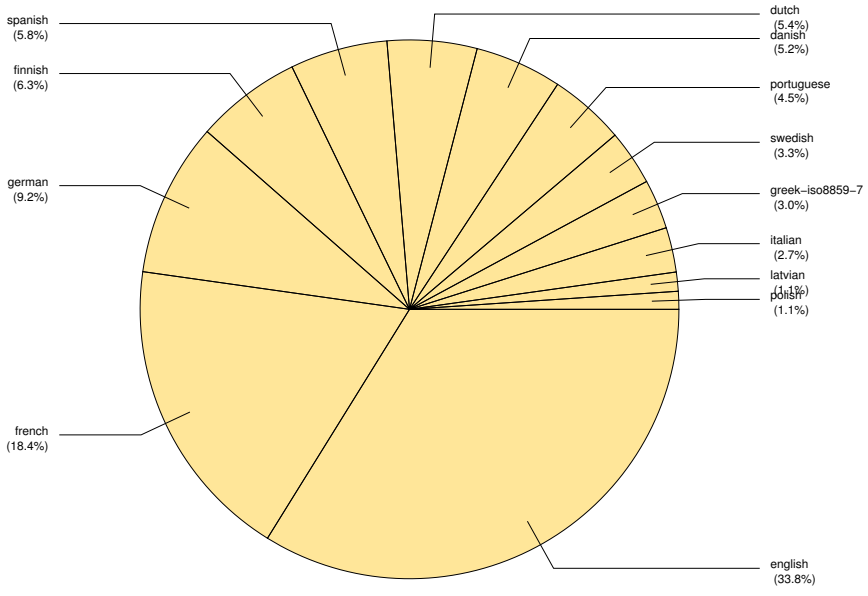


Fig. 1. Language distribution in the `.eu.int` domain.

Table 5. Salient properties of the EuroGOV link structure.

Domain	Number of links		Avg # links/page	
	True	Realized	True	Realized
at	438,591	367,590	43.5759	36.5216
be	729,597	340,415	10.5703	4.93191
cy	42,407	33,231	21.5046	16.8514
cz	10,602,034	6,286,711	32.6723	19.3738
de	15,890,499	2,881,943	35.7179	6.47789
dk	33,572	26,158	15.6586	12.2006
ee	291,709	180,085	17.3968	10.7398
es	318,696	218,448	9.0621	6.21156
eu.int	11,754,603	7,402,189	31.3886	19.7663
fi	17,505,000	3,881,257	26.4602	5.86683
fr	6,101,468	5,080,196	38.9990	32.4713
gr	7,973	6,007	26.3135	19.8251
hu	14,412,108	5,345,513	43.5645	16.1583
ie	397,159	279,833	31.1400	21.9408
it	2,435,376	2,048,472	27.1091	22.8024
lt	161,601	87,511	15.0117	8.12922
lu	186,984	146,270	21.9439	17.1658
lv	9,325,789	5,547,302	29.3807	17.4767
mt	273,873	215,417	19.5749	15.3968
nl	7,087,202	3,636,065	47.2635	24.2484
pl	1,632,655	1,187,235	24.4092	17.7499
pt	9,046,688	5,613,440	61.3564	38.0714
ru	4,880,064	783,246	46.6282	7.48379
se	4,766,234	1,280,677	46.5148	12.4984
si	213,239	152,137	17.1497	12.2356
sk	1,167,119	892,326	20.1155	15.3794
uk	1,847,259	1,286,001	27.8407	19.3818
Avg	4,501,833	2,044,655	29.1971	16.9391
Total	121,549,499	55,205,675		

limitation of the document collection. However, this sort of analysis is important for evaluating whether the collection is a reasonable representative of a realistic web.

- *Empty pages*: The collection contains over 70,000 empty documents. It is not clear why this error ocurred, but it should be avoided in future versions of the collection.
- *Rich document types*: In the EuroGOV collection, document types such as PDF and DOC files appear in the collection in the same format as they were crawled, i.e., their text is not extracted. Furthermore, large documents are truncated to avoid the collection growing too big. A truncated PDF or DOC file does not go down well with several off-the-shelf document parsers. From a web document collection perspective, this is a realistic and interesting scenario. From the perspective of a cross-lingual retrieval collection this

scenario might, however, be less desirable since participants might spend too much time on these issues rather than focusing on the multi-lingual aspects of the task.
- *Character Encodings*: Character encoding is very varied in the European Web, especially for non-latin languages and for extended character sets. Added to that, the information about it in the metadata HTTP header is often wrong, because it is automatically produced and people do not know or care to set it right. Again, this adds to the realism of a cross-lingual web document collection, but also requires considerable effort from participants more interested in the multi-lingual aspects of the collection.

Despite its limitations EuroGOV is very suitable for the initial exploration of cross-lingual web search.

The EuroGOV collection is available for WebCLEF participants, but also as a resource for researchers in fields like natural language processing, information retrieval, or document understanding. Details on how to obtain the EuroGOV collection are on the WebCLEF website [5].

5 Acknowledgments

Thanks to Craig Macdonald for advise at various stages during the creation of EuroGOV, to the Melange group for sharing their normalized links with us, and to Krisztian Balog for help with exploring EuroGOV's link structure.

The building of EuroGOV was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.-069.006, 640.001.501, and 640.002.501.

Bibliography

[1] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[2] .GOV. TREC Web Corpus: .GOV, 2006. URL: `http://es.csiro.au/TRECWeb/govinfo.html`.

[3] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Overview of WebCLEF 2005. In *This Volume*, 2006.

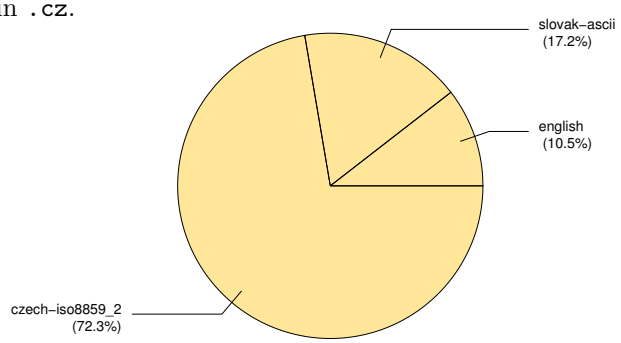
[4] TextCat. Language identification tool, 2006. URL: `http://odur.let.rug.nl/~vannoord/TextCat/`.

[5] WebCLEF. Cross-lingual web retrieval, 2006. URL: `http://ilps.science.uva.nl/WebCLEF/`.

A Language Distributions in EuroGOV

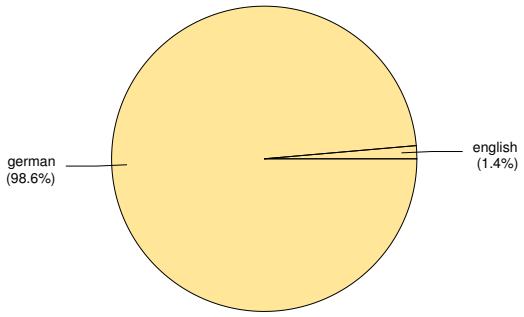
Czech Republic Top-level domain `.cz`.

Domain `.cz`.	
Language	Percentage
czech-iso8859_2	71.71%
slovak-ascii	17.03%
english	10.41%



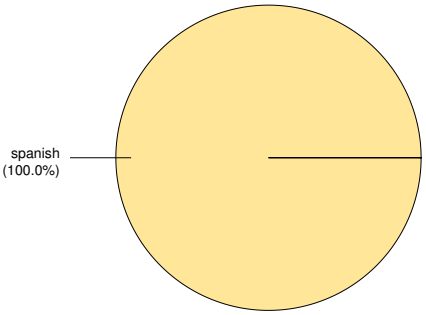
Germany Top-level domain `.de`.

Domain `.de`.	
Language	Percentage
german	97.70%
english	1.37%
french	0.74%



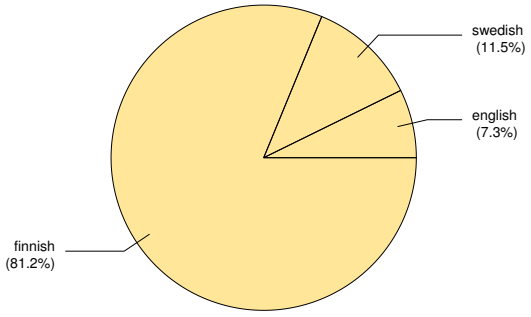
Spain Top-level domain `.es`.

Domain `.es`.	
Language	Percentage
spanish	97.20%
english	0.96%
latvian	0.91%
french	0.86%



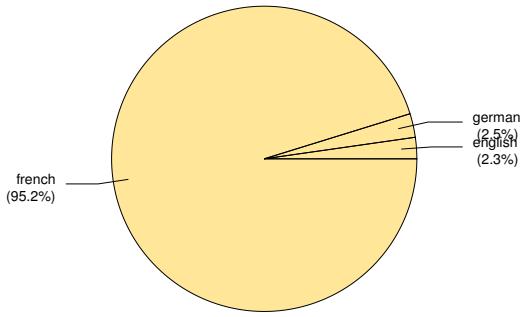
Finland Top-level domain `.fi`.

Domain `.fi`.	
Language	Percentage
finnish	81.15%
swedish	11.52%
english	7.26%



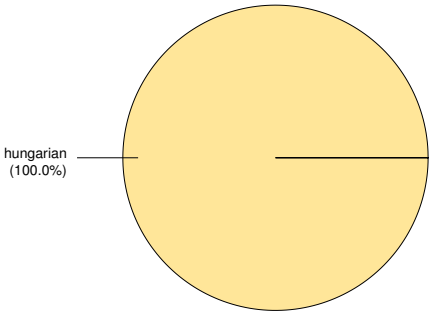
France Top-level domain `.fr`.

Domain `.fr`.	
Language	Percentage
french	94.25%
german	2.49%
english	2.24%
spanish	0.81%



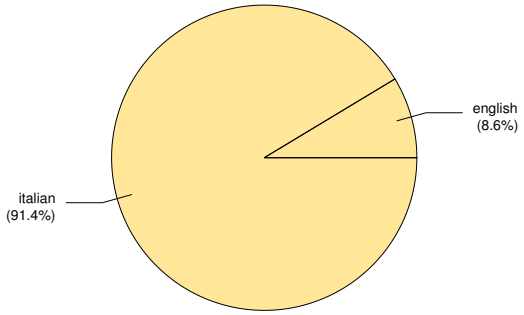
Hungary Top-level domain `.hu`.

Domain `.hu`.	
Language	Percentage
hungarian	99.60%
english	0.31%



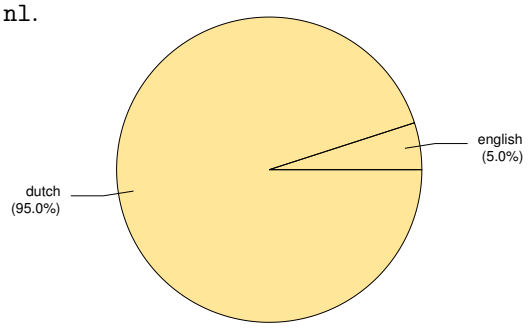
Italy Top-level domain `.it`.

Domain `.it`.	
Language	Percentage
italian	90.15%
english	8.52%
french	0.89%



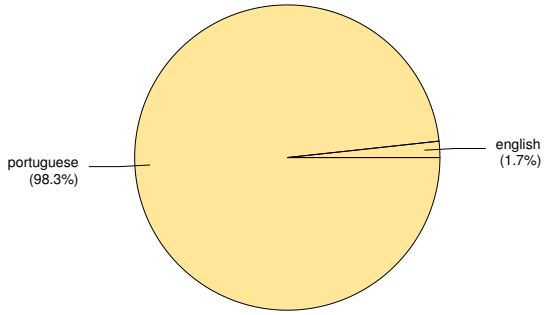
The Netherlands Top-level domain `.nl`.

Domain `.nl`.	
Language	Percentage
dutch	94.39%
english	4.94%



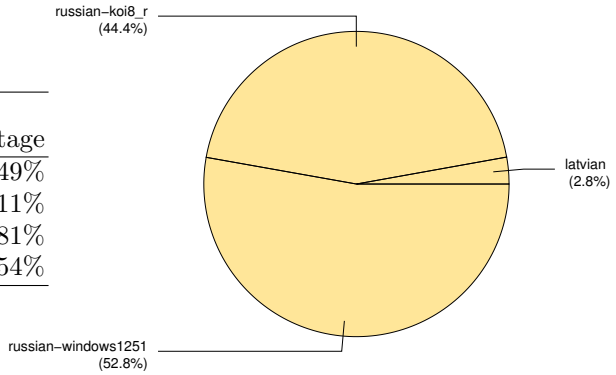
Portugal Top-level domain `.pt`.

Domain `.pt`.	
Language	Percentage
portuguese	98.13%
english	1.72%



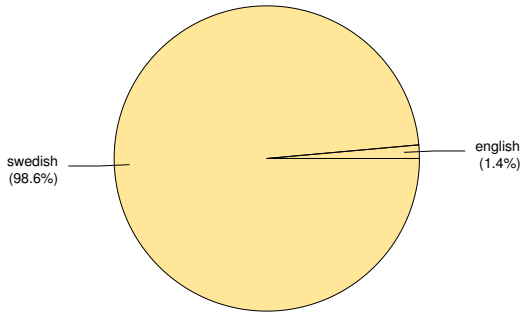
Russia Top-level domain `.ru`.

Domain `.ru`.	
Language	Percentage
russian-windows1251	52.49%
russian-koi8_r	44.11%
latvian	2.81%
english	0.54%



Sweden Top-level domain `.se`.

Domain `.se`.	
Language	Percentage
swedish	98.45%
english	1.42%



United Kingdom Top-level domain `.uk`.

Domain `.uk`.	
Language	Percentage
english	99.05%
welsh	0.47%