Language Richness of the Web

Martin Majliš, Zdeněk Žabokrtský
{majlis, zabokrtsky}@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

W2C Corpus – http://ufal.mff.cuni.cz/~majlis/w2c/

Motivation

There was no multilingual corpus containing many languages with enough data for various NLP applications.

We decided to build a web corpus containing at least 100 languages with a minimum of 10MB of text for each language.

This language data resource can be of use particularly to researchers specialized in development of multilingual technologies.

Building a Web Corpus

Can be divided into several steps as follows:

- retrieving metadata
- building an initial corpus
- generating word tuples
- using them as search queries
- downloading found pages
- removing boilerplate code
- identifying language
- removing duplicate content
- evaluating text quality

Metadata

Combination of multiple sources: SIL International, Wikipedia, and Ethnologue.

These metadata are now accessible through RESTful API.

http://ufal.mff.cuni.cz/~majlis/w2c/api/

Language Identification

YALI, our language identifier, uses a scoring function based on byte 4-grams.

http://ufal.mff.cuni.cz/~majlis/yali/

Duplicity Reduction

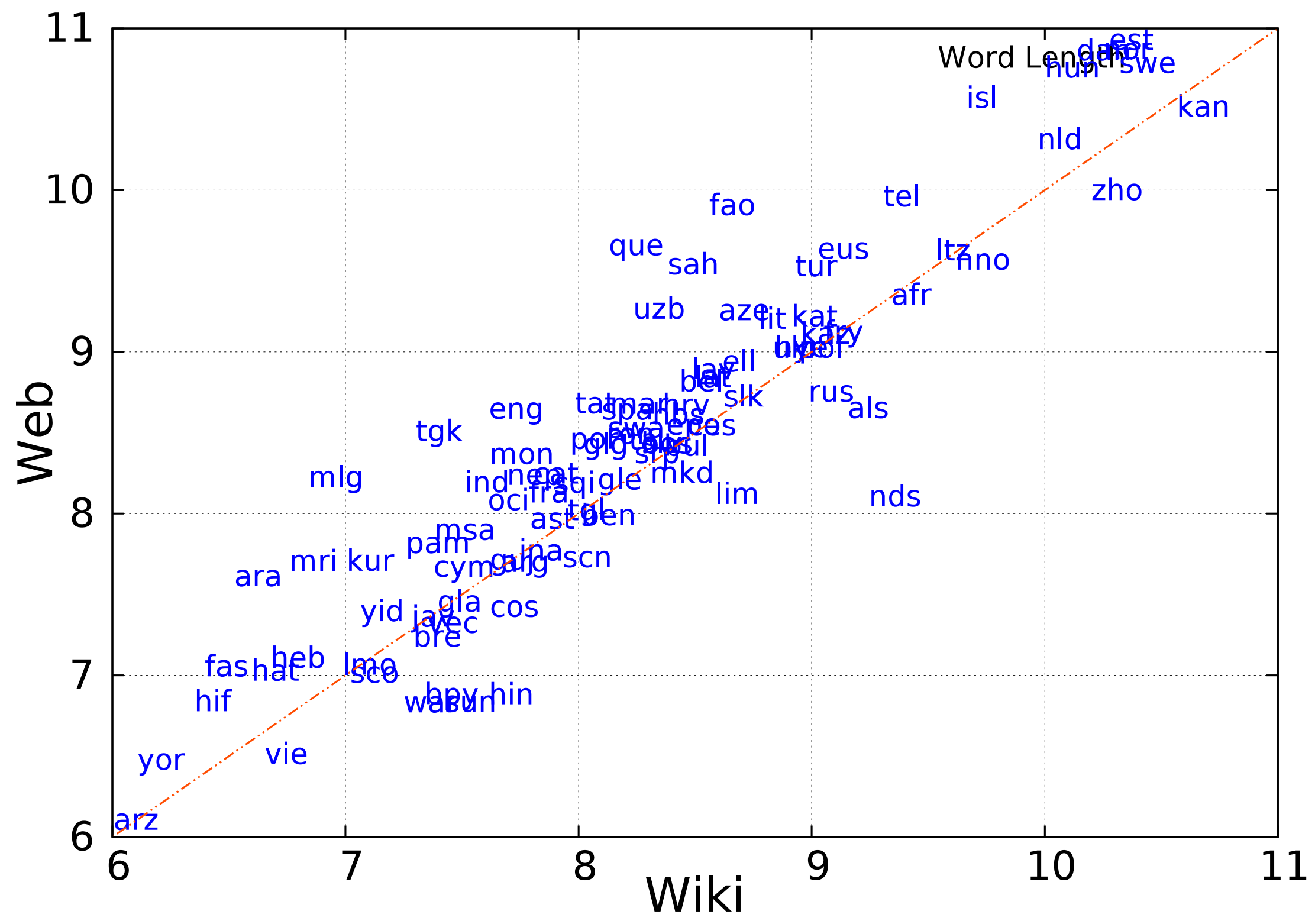
Duplicity sources: spam, common passages and incorrectly detected boilerplate code. Only unique paragraphs are preserved.

Quality Evaluation

We considered Wikipedia a reliable source. Web corpus was compared with Wiki corpus. A difference in certain text property may point to a language for which suspicious material was collected.

Average Word Length

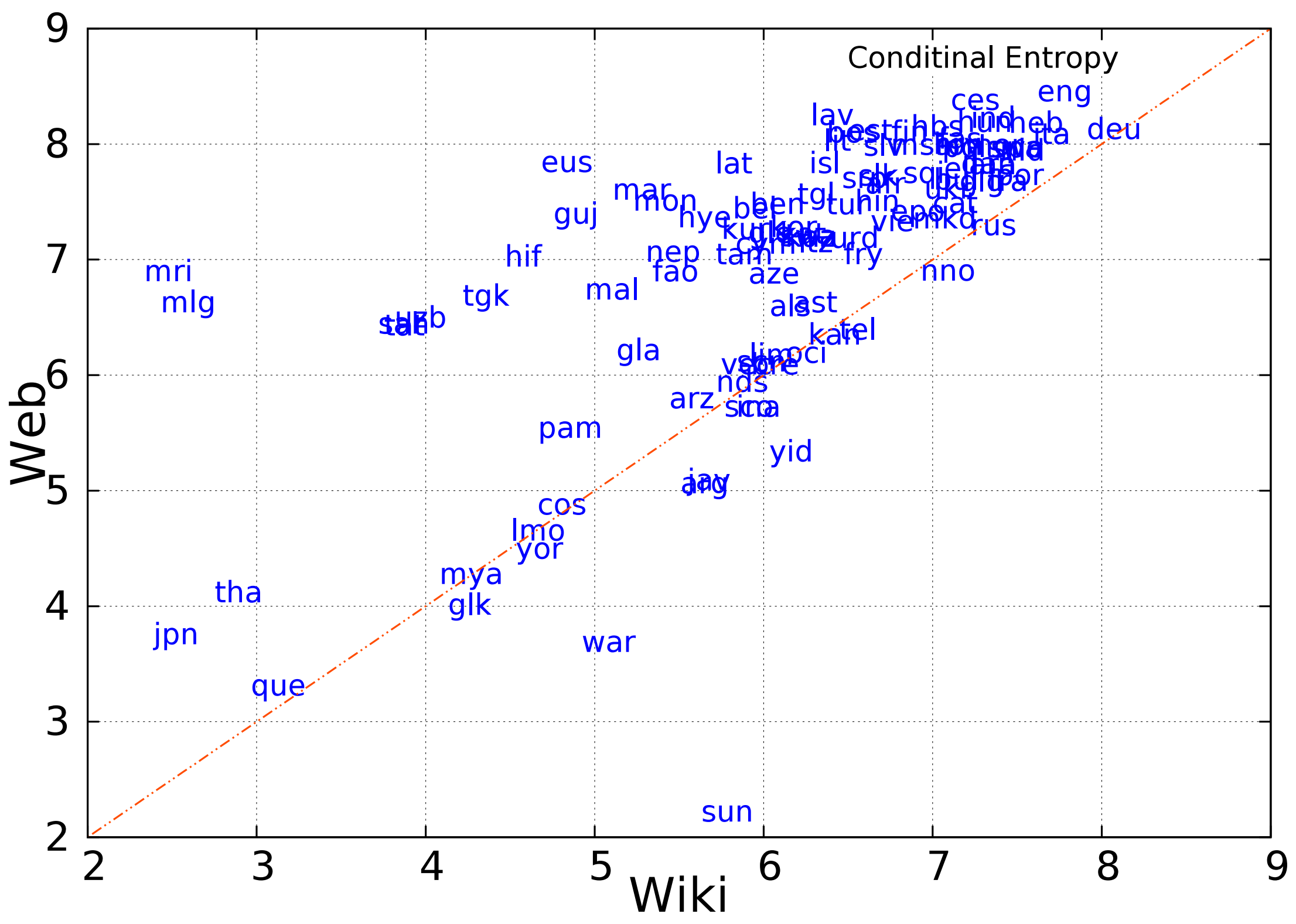
May reveal problems caused by HTML parsing. Average Wiki / Web ratio is 0.973.



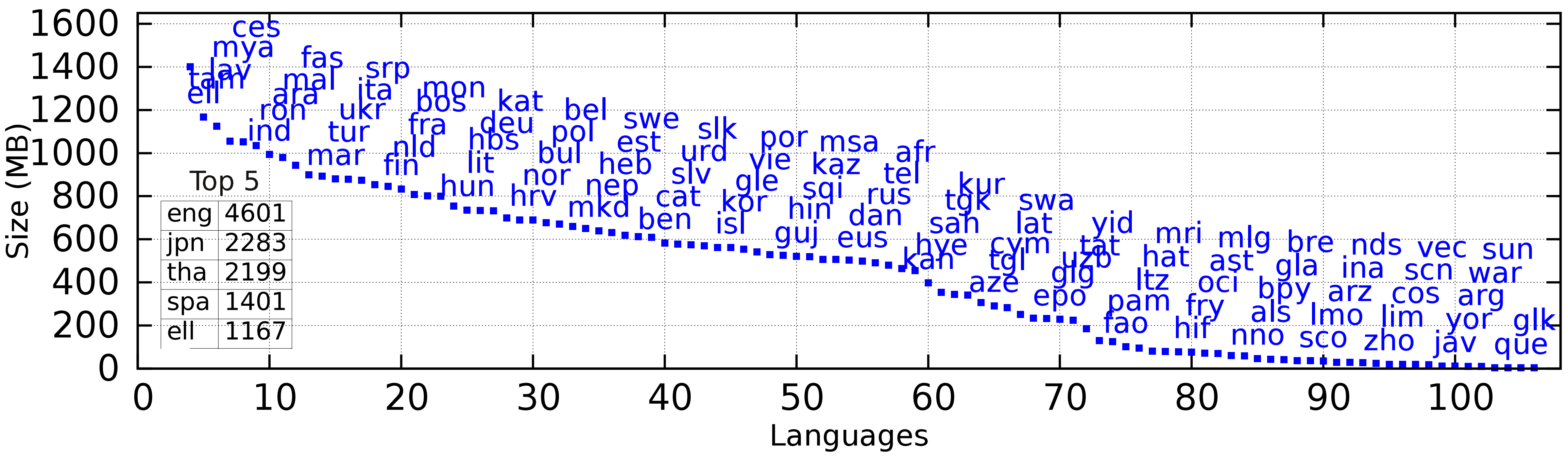
Conditional Entropy

Conditional entropy of the word distribution given the previous word.

Average Wiki / Web ratio is 0.887.



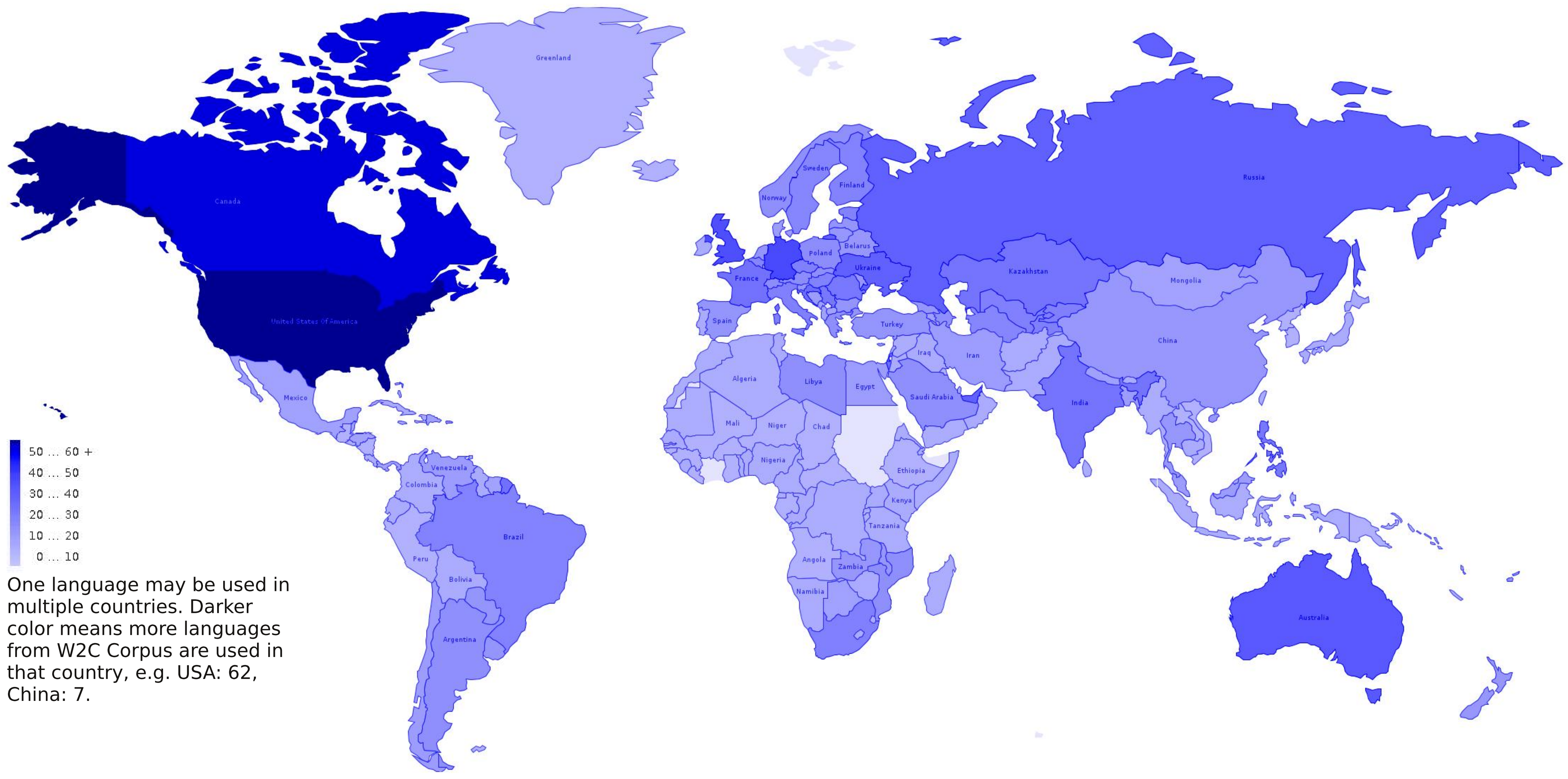
W2C Web Corpus



eng	4601
jpn	2283
tha	2199
spa	1401
ell	1167

World Coverage

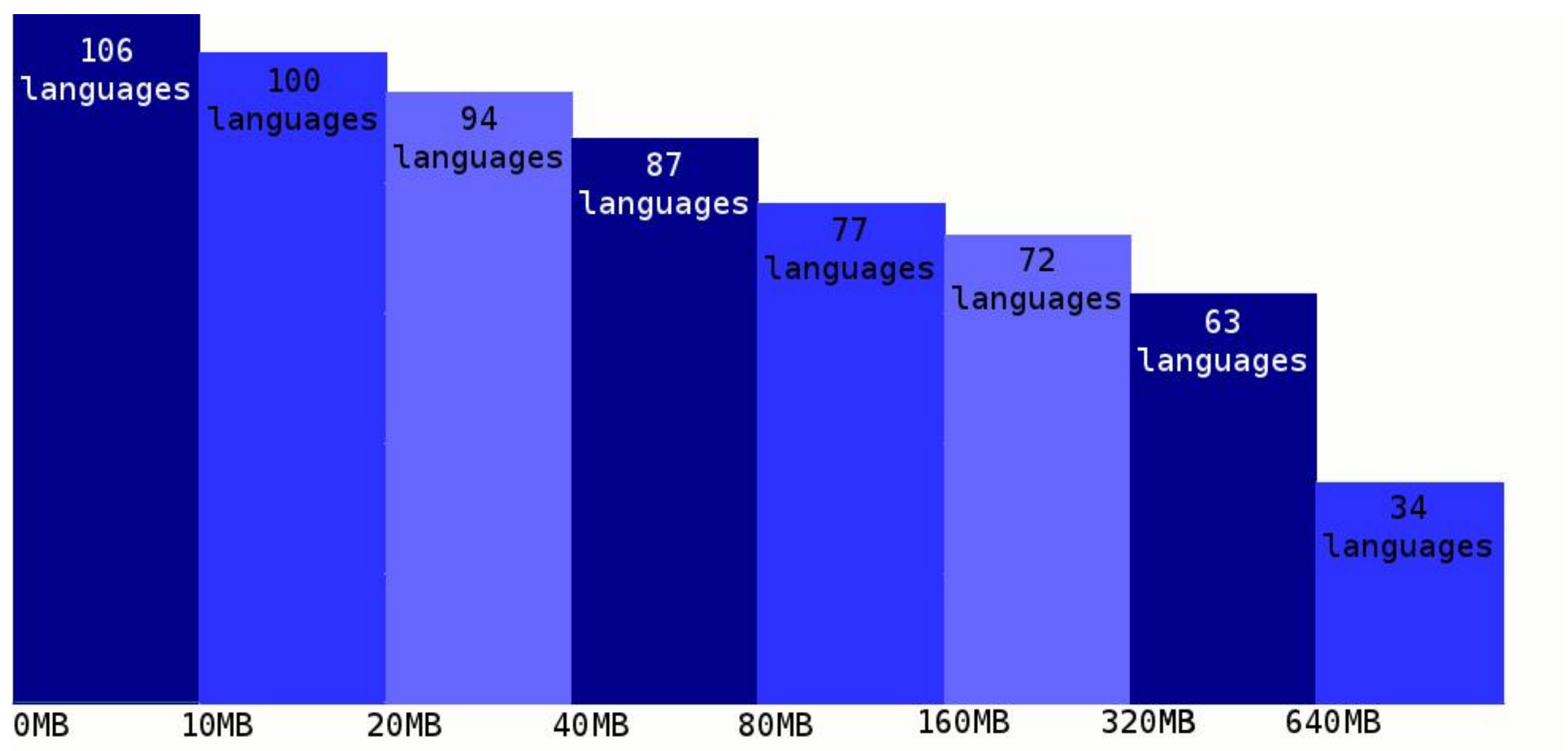
States covered by languages included in W2C Corpus according to Etnologue



One language may be used in multiple countries. Darker color means more languages from W2C Corpus are used in that country, e.g. USA: 62, China: 7.

Corpus Size

- Languages: 106
- Number of URLs downloaded: 103,886,418
- Raw crawl size: 4,554.6 GB
- Raw text size: 131.3 GB
- Unique text size: 54.7 GB



Related Corpora

Number of languages and amount of text in MB.

Corpus	#Lang	Median	Mean	Total
Corpus Factory	8	102.0	85.5	684
Crúbadán 1.0	487	0.068	1.6	769
Crúbadán 2.0	1023	0.127	1.5	1556
I-X	3	126.0	136.0	409
WaCky	3	1500.0	1592.0	4778
W2C Wiki Corpus	**106**	**1.985**	**6.8**	**725**
W2C Web Corpus	**106**	**13.725**	**46.8**	**4961**

- Corpus Factory - Kilgarriff et al. (2010)
- Crúbadán 1.0 - Scannell (2007)
- Crúbadán 2.0 - Scannell (2011)
- I-X – Sharoff (2006)
- WaCky - Baroni et al. (2009)

Languages

The Web and Wiki columns represent text size in MB.

Name	ISO	Web	Wiki
Afrikaans	afr	455	28
Albanian	sqi	507	39
Arabic	ara	943	183
Aragonese	arg	10	16
Armenian	hye	353	22
Asturian	ast	60	12
Azerbaijani	aze	291	61
Basque	eus	499	81
Belarusian	bel	650	46
Bengali	ben	583	51
Bishnupriya	bpy	42	27
Bosnian	bos	799	33
Breton	bre	37	19
Bulgarian	bul	670	169
Burmese	mya	1052	51
Catalan	cat	578	134
Chinese	zho	20	164
Corsican	cos	20	1
Croatian	hrv	690	98
Czech	ces	1035	120
Danish	dan	491	84
Dutch	nld	808	145
Egyptian A	arz	29	9
English	eng	4601	429
Esperanto	epo	229	64
Estonian	est	612	71
Faroese	fao	102	2
Fiji Hindi	hif	77	0
Finnish	fin	833	127
French	fra	802	273
Galician	glg	225	90
Georgian	kat	690	107
German	deu	699	342
Gilaki	glk	4	1
Gujarati	guj	521	64
Haitian	hat	79	6
Hebrew	heb	618	234
Hindi	hin	520	209
Hungarian	hun	736	160
Icelandic	isl	562	25
Indonesian	ind	993	95
Interlingua	ina	27	3
Irish	gle	541	12
Italian	ita	854	211
Japanese	jpn	2283	267
Javanese	jav	12	10
Kannada	kan	398	120
Kazakh	kaz	507	103
Korean	kor	554	138
Kurdish	kur	306	8
Latin	lat	233	19
Latvian	lav	1055	41
Limburgan	lim	20	7
Lithuanian	lit	734	69
Lombard	lmo	29	8
Low German	nds	24	20
Luxembourgish	ltz	81	17
Macedonian	mkd	639	107
Malagasy	mlg	58	11
Malayalam	mal	900	86
Malay	msa	503	72
Maori	mri	78	1
Marathi	mar	880	24
Modern Greek	ell	1167	205
Mongolian	mon	754	14
Nepali	nep	631	23
Norwegian	nor	677	98
Norw Nynorsk	nno	46	61
Occitan	oci	71	12
Pampanga	pam	95	2
Persian	fas	892	137
Polish	pol	660	137
Portuguese	por	525	165
Quechua	que	4	1
Romanian	ron	980	123
Russian	rus	479	350
Scots	sco	35	6
Scottish Gaelic	gla	38	3
Serbian	srp	845	144
Serbo-Croatian	hbs	732	82
Sicilian	scn	19	6
Slovak	slk	562	78
Slovenian	slv	574	73
Spanish	spa	1401	282
Sundanese	sun	4	7
Swahili	swa	232	12
Swedish	swe	610	109
Tagalog	tgl	283	14
Tajik	tgk	342	4
Tamil	tam	1125	148
Tatar	tat	130	10
Telugu	tel	465	130
Thai	tha	2199	228
Tosk Albanian	als	43	16
Turkish	tur	879	107
Ukrainian	ukr	873	214
Urdu	urd	569	25
Uzbek	uzb	185	3
Venetian	vec	13	4
Vietnamese	vie	530	136
Waray	war	4	1
Welsh	cym	251	18
Western Frisian	fry	72	19
Yakut	sah	344	4
Yiddish	yid	125	13
Yoruba	yor	10	1

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). The research has been supported by the grant Khresmoi (FP7-ICT-2010-6-257528 of the EU and 7E11042 of the Czech Republic).

LREC 2012, 23-25 May, Istanbul, Turkey