Corpus collection for under-resourced languages with more than one million speakers

Dirk Goldhahn[1], Maciej Sumalvico[1], Uwe Quasthoff[1,2]
Natural Language Processing Group, University of Leipzig, Germany
Department of African Languages, University of South Africa, South Africa
Email: { dgoldhahn, janicki, quasthoff, }@informatik.uni-leipzig.de

Abstract

For only 40 of about 350 languages with more than one million speakers, the situation concerning text resources is comfortable. For the remaining languages, the number of speakers indicates a need for both corpora and tools. This paper describes a corpus collection initiative for these languages. While random Web crawling has serious limitations, native speakers with knowledge of web pages in their language are of invaluable help. The aim is to give interested scholars, language enthusiasts the possibility to contribute to corpus creation or extension by simply entering a URL into the Web Interface. Using a Web portal URLs of interest are collected with the help of the respective communities. A standardized corpus processing chain for daily newspaper corpora creation is adapted to append newly added web pages to an increasing corpus. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available.

Keywords: corpora, under-resourced languages, Web portal, community

1. Introduction

There are about 350 languages with more than one million speakers[1]. For about 40 of them, the situation concerning text resources is comfortable: there are corpora of reasonable size and also tools like POS taggers adapted to these languages. For the remaining languages, the number of speakers indicates a need for both corpora and tools.
The paper describes a corpus collection initiative for these languages. While random Web crawling has serious limitations, native speakers with knowledge of web pages in their language are of invaluable help. The aim is to give interested scholars, language enthusiasts the possibility to contribute to corpus creation or extension by simply entering a URL into the Web Interface. Using a Web portal URLs of interest are collected with the help of the respective communities. A standardized corpus processing chain for daily newspaper corpora creation is adapted to append newly added web pages to an increasing corpus. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available directly and as part of the Leipzig Corpora Collection.

2. Crawling strategies and limitations

Random web crawling for smaller languages has several limitations. They are among others related to aspects like the relatively small amount of web pages, the inadequate link structure and the ranking on search engines.

2.1. General crawling problems

The following crawling problem applies to all the following strategies: Due to technical limitations many crawlers cannot follow all links. So-called JavaScript links require the execution of JavaScript code by the crawler which often produces errors. So, if a website heavily uses JavaScript links and there are no other links pointing to special pages (coming from another website, for instance), then all but the main page might be excluded from crawling. At the time of writing (autumn 2015), only the Google crawler is assumed to be able to follow all JavaScript links.
This is a problem especially if the linking density is low, i.e. if the number of static links is not enough to reach most pages. For smaller languages often the web community is in an initial state so complete crawling is difficult.

2.2. Random Web Crawling

The naive algorithm would crawl as much as possible and classify the web pages by language. Modern crawlers like Heritrix[2] [Mohr, 2004] are able to crawl hundreds of millions of pages on commodity hardware, so this should be no problem. The results of such a crawling are even available for direct download: The Common Crawl[3] collected more than 1.8 billion web pages, but the focus is on certain TLDs (Top Level Domains) and on the 40 most prominent languages. Other languages are underrepresented[4].
Language identification in random collections also is more difficult than in restricted collections (see next subsection): The number of pages to be identified in the big collection is comparatively very small, so false positives are a problem. False positives can come from apparently similar languages spoken elsewhere in the

1 http://www.ethnologue.com/

2 https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
3 https://commoncrawl.org/
4 https://docs.google.com/file/d/1_9698uglerxB9nAglvaHkEgU-iZNm1TvVGuCW7245-WGvZq47teNpb_uL5N9/edit

world and from non-text junk documents containing some words of the language, but no actual meaningful text.

2.3. Crawling TLDs

In the simplest case a language is spoken only within one country. Then the natural approach is first to crawl all Web pages of the corresponding top-level domain (TLD) and then to extract the pages in the desired language. The same approach works if a language is spoken in a small number of different countries. Language identification usually also works well [McNamee, 2005] because in most cases the language under consideration is easy to be distinguished from the other languages spoken in the corresponding country.

But for some languages this approach is not productive. Especially in the case of non-official minority languages the community of speakers sometimes prefers another TLD (like .com) instead of the own country's TLD. And in the case of the .com domain, language identification is no more reliable as described above.

2.4. BootCaT approach

In an approach similar to Baroni [2004], frequent terms of a language are combined to form search queries for engines such as Google and to retrieve the resulting URLs as basis for later crawling. As a requirement a small set of frequent terms is needed for each language. Documents such as the Universal Declaration of Human Rights (UDHR), which is available in more than 350 languages, are one possible resource to build such word lists. For an average language, the UDHR contains about 2,000 running words which is still suitable for the task described. An alternative source for word lists are Watchtower documents, which are also available online for about 200 languages[5].

Based on these resources, lists of word tuples of three to five high frequent words are generated. These tuples are then used to query Web search engines and to collect the retrieved URLs. In a next step these Web sites are downloaded and processed further.

Unfortunately, there are certain limitations to the use of this approach for lesser resourced languages. Typically there is a very limited number of resources in the language in question. Since the communities in these languages are small Web sites of interest are typically sparsely linked. Therefore a low page rank score occurs and as a result the Web sites are typically ranked badly on search engines. Experience when using this approach have shown that high ranked results are most likely English Web pages containing few words of the language in question or the short translation of text passages in that language. When checking the Web page for its common language, such URLs are dismissed since English will be detected.

5 https://www.jw.org/

3. Collection method

In this section we propose an alternative method to collect textual resources from the Web for under-resourced languages. The basis of this approach are native speakers with knowledge of Web sites in their respective language. The aim is to give interested scholars, language enthusiasts the possibility to contribute to corpus creation or extension by simply entering a URL into the Web Interface. In order to facilitate the gathering of URLs for a large number of languages a Web portal is currently being developed[6]. Using a basic input mask (see Figure 1) users can easily add URLs together with information on the languages present on this web resource. All additional input fields are optional.

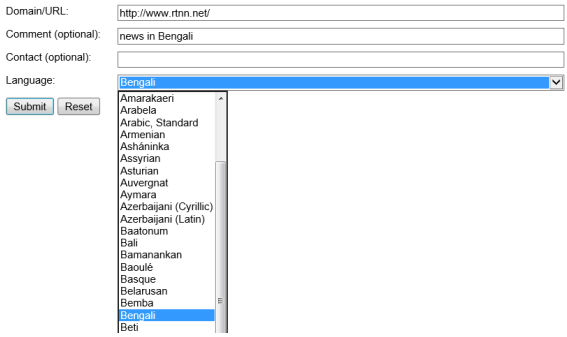

Figure 1: Input mask for URLs of language resources on the Web portal.

The data entered can be viewed online (see Figure 2) and is stored in a local mysql-storage engine where it serves as input for deeper analysis and corpus creation. In a first step parts of a domain are downloaded using Heritrix the crawler of the Internet Archive. Using statistical language identification [Pollmächer, 2012] the languages present in the respective domain are determined. As a data basis for comparison web corpora or documents from sources such as Universal Declaration of Human Rights or Watchtower for several hundred languages are utilized. These sources can include multiple entries for languages using more than one script, enabling the system to create respective corpora.

Bengali (ben)

URL/Domain	Comment
http://www.anandabazar.com/	mostly Bengali
http://www.atnbangla.tv/	TV station
http://www.dainikdestiny.com/	
http://www.jugantor.com	
http://www.rtnn.net/bangla/	news in Bengali
http://www.thedailysangbad.com/	news Website

Figure 2: List view of entries for the Bengali language on the Web portal.

In case the desired language is at least partly present in the documents crawled, in the next step the whole domain is being downloaded. Results of this process are then

6 Available soon at http://small-languages.informatik.uni-leipzig.de

processed utilizing a standardized corpus processing chain for daily newspaper corpora creation[7] which has been adapted to append newly added web pages to an increasing corpus for each language. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available.
The Web portal and the adapted processing chain are currently in development. Both will be finished in March 2016.

4. Corpus Processing Chain

We apply a standardized, language-independent pipeline for building corpora from raw data. For details on this processing chain please see Goldhahn (2012). We use own tools for extracting raw text from WARC files (Heritrix output) and HTML pages. Then we apply statistical language identification on document basis. Further processing steps are: sentence segmentation, removal of ill-formed sentences based on handwritten regular expressions, language identification on sentence basis, duplicate sentence removal (a removal of near duplicates such as boilerplates is currently in development), tokenization and word co-occurrence calculation. Finally, the corpora are stored as MySQL databases with a standardized schema. In addition to the basic workflow, additional (possibly language-specific) tools can be applied to some corpora, like POS-tagging, which results in additional database tables.
A couple of technical issues must be taken care of in multilingual processing. As we are using UTF-8 as the sole encoding, proper conversion must be guaranteed at the preprocessing step (HTML/WARC → text). The sentence separator is a rule based tool, which requires a list of sentence-terminating characters. It is important to include such characters for all expected languages and writing systems. Pairs of characters that look similar, but are encoded differently, like Latin semicolon (U+003B) and Greek question mark (U+037E), need special attention. For language segmentation, lists of around 1K most frequent words for each language need to be supplied.
Tests on various input data have shown that our processing chain handles data volumes of up to 200 million sentences. For corpora of 100K - 1M sentences, the running times are typically less than an hour.

5. Conclusion

This paper describes a corpus collection initiative for lesser resourced languages, enabling scholars or language enthusiasts to create and extend corpora for these languages by simply entering a URL. Using this Web portal URLs of interest are collected with the help of the respective communities. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available.

7 http://wortschatz.uni-leipzig.de/wort-des-tages/

6. Bibliographical References

Baroni, M., & Bernardini, S. (2004, May). BootCaT: Bootstrapping Corpora and Terms from the Web. In LREC.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In LREC (pp. 759-765).

McNamee, P. (2005). Language identification: a solved problem suitable for undergraduate instruction. Journal of Computing Sciences in Colleges, 20(3), 94-101.

Mohr, G., Stack, M., Rnitovic, I., Avery, D., & Kimpton, M. (2004, July). Introduction to heritrix. In 4th International Web Archiving Workshop.

Pollmächer, J. (2011). Separierung mit FindLinks gecrawlter Texte nach Sprachen. Bachelor Thesis, University of Leipzig.

7. Appendix: Language Lists

The following lists are collected as follows: The languages with more than 1 million of speakers (according to Ethnologue[8]) are divided into two parts: The Leipzig Corpora Collection[9] [Goldhahn et al., 2012] is used to identify under-resourced languages. For simplicity, a language is called under-resourced if there are less than 1 million sentences in the corpora of this collection. The following two tables show both the under-resourced languages and the well-resourced languages with more than 1 million sentences in the collection.

Language	*Code*	*Country*
Dari	prs	Afghanistan
Hazaragi	haz	Afghanistan
Albanian	sqi	Albania
Kabyle	kab	Algeria
Tachawit	shy	Algeria
Kimbundu	kmb	Angola
Umbundu	umb	Angola
Armenian	hye	Armenia
Bengali	ben	Bangladesh
Chittagonian	ctg	Bangladesh
Rangpuri	rkt	Bangladesh
Sylheti	syl	Bangladesh
Vlaams	vls	Belgium
Fon	fon	Benin
Aymara	aym	Bolivia
Quechua	que	Bolivia, Peru
Bosnian	bos	Bosnia and Herzegovina
Hunsrik	hrx	Brazil
Jula	dyu	Burkina Faso

8 http://www.ethnologue.com/
9 http://corpora.informatik.uni-leipzig.de/

Mòoré	mos	Burkina Faso
Rundi	run	Burundi
Central Khmer	khm	Cambodia
Fulah	ful	Cameroon
Kabuverdianu	kea	Cape Verde Islands
Bouyei	pcc	China
Chuanqiandian Cluster Miao	cqd	China
Gan Chinese	gan	China
Hmong	hmn	China
Hmong Daw	mww	China
Khams Tibetan	khg	China
Min Dong Chinese	cdo	China
Min Nan Chinese	nan	China
Northern Qiandong Miao	hea	China
Nuosu	iii	China
Southern Dong	kmc	China
Tibetan	bod	China
Uyghur	uig	China
Zhuang	zha	China
Kituba	mkw	Congo
Baoulé	bci	Côte d'Ivoire
Dan	dnj	Côte d'Ivoire
Alur	alz	Dem. Rep. of Congo
Chokwe	cjk	Dem. Rep. of Congo
Kituba	ktu	Dem. Rep. of Congo
Kongo	kon	Dem. Rep. of Congo
Koongo	kng	Dem. Rep. of Congo
Lingala	lin	Dem. Rep. of Congo
Luba-Kasai	lua	Dem. Rep. of Congo
Luba-Katanga	lub	Dem. Rep. of Congo
Ngbaka	nga	Dem. Rep. of Congo
Songe	sop	Dem. Rep. of Congo
Yombe	yom	Dem. Rep. of Congo
Zande	zne	Dem. Rep. of Congo
Tigré	tig	Eritrea
Afar	aar	Ethiopia
Amharic	amh	Ethiopia
Gamo	gmv	Ethiopia
Hadiyya	hdy	Ethiopia
Oromo	orm	Ethiopia
Sidamo	sid	Ethiopia
Wolaytta	wal	Ethiopia
Tigrigna	tir	Ethiopia, Eritrea
Occitan	oci	France
Abron	abr	Ghana
Akan	aka	Ghana
Éwé	ewe	Ghana

Pontic	pnt	Greece
Quiché	quc	Guatemala
Eastern Maninkakan	emk	Guinea
Fang	fan	Guinea
Kpelle	kpe	Guinea
Pular	fuf	Guinea
Susu	sus	Guinea
Haitian	hat	Haiti
Ahirani	ahr	India
Assamese	asm	India
Awadhi	awa	India
Bagheli	bfy	India
Bhili	bhb	India
Bhojpuri	bho	India
Bodo	brx	India
Bundeli	bns	India
Chhattisgarhi	hne	India
Deccan	dcc	India
Dhundari	dhd	India
Dogri	doi	India
Garhwali	gbm	India
Garo	grt	India
Goan Konkani	gom	India
Godwari	gdx	India
Gondi	gon	India
Haryanvi	bgc	India
Ho	hoc	India
Kanauji	bjj	India
Kangri	xnr	India
Kannada	kan	India
Kashmiri	kas	India
Konkani	knn	India
Kumaoni	kfy	India
Kurux	kru	India
Lambadi	lmn	India
Magahi	mag	India
Mahasu Pahari	bfz	India
Maithili	mai	India
Malayalam	mal	India
Marwari	mwr	India
Meitei	mni	India
Mina	myi	India
Mundari	unr	India
Nimadi	noe	India
Oriya	ori	India
Rajasthani	raj	India
Sadri	sck	India
Santali	sat	India
Shekhawati	swv	India

Surgujia	sgj	India
Surjapuri	sjp	India
Tamil	tam	India
Telugu	tel	India
Tulu	tcy	India
Varhadi-Nagpuri	vah	India
Vasavi	vas	India
Jambi Malay	jax	Indonesia
Bali	ban	Indonesia (Java and Bali)
Betawi	bew	Indonesia (Java and Bali)
Javanese	jav	Indonesia (Java and Bali)
Madura	mad	Indonesia (Java and Bali)
Sunda	sun	Indonesia (Java and Bali)
Banjar	bjn	Indonesia (Kalimantan)
Sasak	sas	Indonesia (Nusa Tenggara)
Bugis	bug	Indonesia (Sulawesi)
Gorontalo	gor	Indonesia (Sulawesi)
Makasar	mak	Indonesia (Sulawesi)
Aceh	ace	Indonesia (Sumatra)
Batak Dairi	btd	Indonesia (Sumatra)
Batak Mandailing	btm	Indonesia (Sumatra)
Batak Simalungun	bts	Indonesia (Sumatra)
Batak Toba	bbc	Indonesia (Sumatra)
Minangkabau	min	Indonesia (Sumatra)
Musi	mui	Indonesia (Sumatra)
Bakhtiari	bqi	Iran
Domari	rmt	Iran
Gilaki	glk	Iran
Iranian Persian	pes	Iran
Kashkay	qxq	Iran
Laki	lki	Iran
Mazanderani	mzn	Iran
Northern Luri	lrc	Iran
Southern Kurdish	sdh	Iran
Central Kurdish	ckb	Iraq
Eastern Yiddish	ydd	Israel
Dholuo	luo	Kenya
Ekegusii	guz	Kenya
Gikuyu	kik	Kenya
Kalenjin	kln	Kenya
Kamba	kam	Kenya
Kimîîru	mer	Kenya
Kipsigis	sgc	Kenya

Lubukusu	bxk	Kenya
Maasai	mas	Kenya
Oluluyia	luy	Kenya
Kurdish	kur	Kurdistan, Iraq, Turkey
Kyrgyz	kir	Kyrgyzstan
Lao	lao	Laos
Macedonian	mkd	Macedonia
Malagasy	mlg	Madagascar
Nyanja	nya	Malawi
Tumbuka	tum	Malawi
Yao	yao	Malawi
Malay	zlm	Malaysia (Peninsular)
Bamanankan	bam	Mali
Maasina Fulfulde	ffm	Mali
Soninke	snk	Mali
Hassaniyya	mey	Mauritania
Halh Mongolian	khk	Mongolia
Central Atlas Tamazight	tzm	Morocco
Tachelhit	shi	Morocco
Tarifit	rif	Morocco
Lomwe	ngl	Mozambique
Makhuwa	vmw	Mozambique
Makhuwa-Meetto	mgh	Mozambique
Sena	seh	Mozambique
Tswa	tsc	Mozambique
Burmese	mya	Myanmar
Pwo Eastern Karen	kjp	Myanmar
Rohingya	rhg	Myanmar
S'gaw Karen	ksw	Myanmar
Shan	shn	Myanmar
Ndonga	ndo	Namibia
Eastern Tamang	taj	Nepal
Nepali	nep	Nepal
Limburgish	lim	Netherlands
Tamashek	tmh	Niger
Zarma	dje	Niger
Anaang	anw	Nigeria
Berom	bom	Nigeria
Central Kanuri	knc	Nigeria
Ebira	igb	Nigeria
Edo	bin	Nigeria
Hausa	hau	Nigeria
Ibibio	ibb	Nigeria
Igbo	ibo	Nigeria
Izon	ijc	Nigeria
Kanuri	kau	Nigeria
Nigerian Fulfulde	fuv	Nigeria
Nigerian Pidgin	pcm	Nigeria

Tiv	tiv	Nigeria
Yoruba	yor	Nigeria
Norwegian	nor	Norway
Baluchi	bal	Pakistan
Brahui	brh	Pakistan
Eastern Balochi	bgp	Pakistan
Lahnda	lah	Pakistan
Northern Hindko	hno	Pakistan
Pahari-Potwari	phr	Pakistan
Seraiki	skr	Pakistan
Sindhi	snd	Pakistan
Southern Balochi	bcc	Pakistan
Western Balochi	bgn	Pakistan
Western Panjabi	pnb	Pakistan
Guarani	grn	Paraguay, Bolivia
Bikol	bik	Philippines
Cebuano	ceb	Philippines
Central Bikol	bcl	Philippines
Filipino	fil	Philippines
Hiligaynon	hil	Philippines
Ilocano	ilo	Philippines
Maguindanao	mdh	Philippines
Pampangan	pam	Philippines
Pangasinan	pag	Philippines
Tagalog	tgl	Philippines
Tausug	tsg	Philippines
Waray-Waray	war	Philippines
Romany	rom	Romania
Bashkort	bak	Russian Federation
Chechen	che	Russian Federation
Chuvash	chv	Russian Federation
Kabardian	kbd	Russian Federation
Tatar	tat	Russian Federation
Rwanda	kin	Rwanda
Mandingo	man	Senegal
Mandinka	mnk	Senegal
Pulaar	fuc	Senegal
Serer-Sine	srr	Senegal
Wolof	wol	Senegal
Mende	men	Sierra Leone
Themne	tem	Sierra Leone
Somali	som	Somalia
Northern Sotho	nso	South Africa
Southern Ndebele	nbl	South Africa
Tsonga	tso	South Africa
Venda	ven	South Africa
Xhosa	xho	South Africa
Zulu	zul	South Africa
Tswana	tsn	South Africa, Botswana
Southern Sotho	sot	South Africa, Lesotho
Swati	ssw	South Africa, Swaziland
Galician	glg	Spain
Sinhala	sin	Sri Lanka
Bedawiyet	bej	Sudan
Dinka	din	Sudan
Tajiki	tgk	Tajikistan
Gogo	gog	Tanzania
Haya	hay	Tanzania
Makonde	kde	Tanzania
Nyakyusa-Ngonde	nyy	Tanzania
Sukuma	suk	Tanzania
Swahili	swa	Tanzania
Northern Khmer	kxm	Thailand
Thai	tha	Thailand
Malay	msa	Thailand, Malaysia
Dimli	diq	Turkey
Zaza	zza	Turkey
Turkmen	tuk	Turkmenistan
Acholi	ach	Uganda
Chiga	cgg	Uganda
Ganda	lug	Uganda
Lango	laj	Uganda
Lugbara	lgg	Uganda
Masaaba	myx	Uganda
Nyankore	nyn	Uganda
Soga	xog	Uganda
Teso	teo	Uganda
Uzbek	uzb	Uzbekistan
Muong	mtq	Viet Nam
Tày	tyz	Viet Nam
Bemba	bem	Zambia
Tonga	toi	Zambia, Zimbabwe
Manyika	mxc	Zimbabwe
Ndau	ndc	Zimbabwe
Ndebele	nde	Zimbabwe
Shona	sna	Zimbabwe

Table 1: Under-resourced languages with more than 1 million speakers and less than 1 million sentences, ordered by country.

Language	*Code*	*Country*
Arabic	ara	various countries
English	eng	various countries
Pushto	pus	Afghanistan, Pakistan
Azerbaijani	aze	Azerbaijan
Belarusan	bel	Belarus
Bulgarian	bul	Bulgaria
Chinese	zho	China
Serbo-Croatian	hbs	Croatia, Serbia, Bosnia and Herzegovina
Czech	ces	Czech Republic
Danish	dan	Danmark
Estonian	est	Estonia
Finnish	fin	Finland
French	fra	France
Georgian	kat	Georgia
Greek	ell	Greece
Hungarian	hun	Hungaria
Gujarati	guj	India
Hindi	hin	India
Marathi	mar	India
Indonesian	ind	Indonesia
Gilaki	glk	Iran
Persian	fas	Iran
Hebrew	heb	Israel
Italian	ita	Italy
Japanese	jpn	Japan
Kazakh	kaz	Kazakhstan
Korean	kor	Korea
Latvian	lav	Latvia
Lithuanian	lit	Lithuania
Mongolian	mon	Mongolia
Dutch	nld	Netherlands
Polish	pol	Poland
Portuguese	por	Portugal
Romanian	ron	Romania
Russian	rus	Russia
Serbian	srp	Serbia
Slovak	slk	Slovakia
Slovene	slv	Slovenia
Afrikaans	afr	South Africa
Catalan-Valencian-Balear	cat	Spain
Spanish	spa	Spain
Tamil	tam	Sri Lanka, India
Swedish	swe	Sweden
Turkish	tur	Turkey
Ukrainian	ukr	Ukraine
Urdu	urd	Urdu
Vietnamese	vie	Vietnam

Table 2: Well-resourced languages with more than 1 million sentences, ordered by country.