CURL – Crawling Under-Resourced Languages

Dear colleagues,
Please help collecting language resources for under-resourced languages! If you provide URLs for Web pages, we will add the extracted texts to the corresponding corpus and provide all data for online access and free download under the Creative Commons License CC-BY.
For the development of Language Technology products, text resources are essential. The NLP Group of the Computer Science department of Leipzig University collects generic text corpora and makes them freely available at http://corpora.uni-leipzig.de. In the CURL project, we concentrate on the collection of texts for languages with more than one million speakers (see next page), for which digital resources should be obtainable. More languages can be added as needed.
The corpora come with the following data:

- sentences in random order (due to copyright restrictions), mostly well-formed (checked using regular expressions, no check of syntax or orthography);
- word lists with frequencies;
- word co-occurrence data: Pairs of words which co-occur significantly often either as immediate neighbors or within the same sentence.

All enlarged corpora can be queried at http://corpora.uni-leipzig.de immediately.

You can help to collect more text in the following way: Please visit http://curl.corpora.uni-leipzig.de/. In the first step, you select your language of interest. In the second step, you enter the URL(s) either in the input field or you upload a text file with one URL per line. Then just press "Process URLs". The URLs provided are the starting points for a crawl of the whole domain which stops either if all pages are crawled or a time limit of several hours is exceeded. The extracted texts will be appended to the existing corpus. If you provide your email address, you will be notified after the processing. You can always find information about the status of your job in the "Process Status" tab.

It would be of great help if you support this text collection project for under-resourced languages.
For questions about this project, contact Dirk Goldhahn, dgoldhahn@informatik.uni-leipzig.de.

Many thanks in advance

Uwe Quasthoff and Dirk Goldhahn

NLP Group,
Dept. Computer Science
University Leipzig

Reference:
Dirk Goldhahn, Maciej Sumalvico und Uwe Quasthoff: *Corpus collection for under-resourced languages with more than one million speakers*. In: *Workshop on Collaboration and Computing for Under-Resourced Languages (CCURL), LREC, Portorož*, 2016

PS: Please feel free to distribute this message to colleagues and students.

List of CURL Languages (2017)

Abron [abr] - Ghana,
Aceh [ace] - Indonesia (Sumatra),
Acholi [ach] - Uganda,
Afar [aar] - Ethiopia,
Ahirani [ahr] - India,
Akan [aka] - Ghana,
Alur [alz] - Dem. Rep. of Congo,
Amharic [amh] - Ethiopia,
Anaang [anw] - Nigeria,
Assamese [asm] - India,
Awadhi [awa] - India,
Aymara [aym] - Bolivia,
Bagheli [bfy] - India,
Bakhtiari [bqi] - Iran,
Bali [ban] - Indonesia (Java and Bali),
Baluchi [bal] - Pakistan,
Bamanankan [bam] - Mali,
Banjar [bjn] - Indonesia (Kalimantan),
Baoulé [bci] - Côte d'Ivoire,
Bashkort [bak] - Russian Federation,
Batak Dairi [btd] - Indonesia (Sumatra),
Batak Mandailing [btm] - Indonesia (Sumatra),
Batak Simalungun [bts] - Indonesia (Sumatra),
Batak Toba [bbc] - Indonesia (Sumatra),
Bedawiyet [bej] - Sudan,
Bemba [bem] - Zambia,
Bengali [ben] - Bangladesh,
Berom [bom] - Nigeria,
Betawi [bew] - Indonesia (Java and Bali),
Bhili [bhb] - India,
Bhojpuri [bho] - India,
Bikol [bik] - Philippines,
Bodo [brx] - India,
Bosnian [bos] - Bosnia and Herzegovina,
Bouyei [pcc] - China,
Brahui [brh] - Pakistan,
Bugis [bug] - Indonesia (Sulawesi),
Bundeli [bns] - India,
Burmese [mya] - Myanmar,
Cebuano [ceb] - Philippines,
Central Atlas Tamazight [tzm] - Morocco,
Central Bikol [bcl] - Philippines,
Central Kanuri [knc] - Nigeria,
Central Khmer [khm] - Cambodia,
Central Kurdish [ckb] - Iraq,
Chechen [che] - Russian Federation,
Chhattisgarhi [hne] - India,
Chiga [cgg] - Uganda,
Chittagonian [ctg] - Bangladesh,
Chokwe [cjk] - Dem. Rep. of Congo,
Chuanqiandian Cluster Miao [cqd] - China,
Chuvash [chv] - Russian Federation,
Dan [dnj] - Côte d'Ivoire,
Dari [prs] - Afghanistan,
Deccan [dcc] - India,
Dholuo [luo] - Kenya,
Dhundari [dhd] - India,
Dimli [diq] - Turkey,
Dinka [din] - Sudan,
Dogri [doi] - India,
Domari [rmt] - Iran,
Eastern Balochi [bgp] - Pakistan,
Eastern Maninkakan [emk] - Guinea,
Eastern Tamang [taj] - Nepal,
Eastern Yiddish [ydd] - Israel,
Ebira [igb] - Nigeria,
Edo [bin] - Nigeria,
Ekegusii [guz] - Kenya,
Éwé [ewe] - Ghana,
Fang [fan] - Guinea,
Filipino [fil] - Philippines,
Fon [fon] - Benin,
Fulah [ful] - Cameroon,
Galician [glg] - Spain,
Gamo [gmv] - Ethiopia,
Gan Chinese [gan] - China,
Ganda [lug] - Uganda,
Garhwali [gbm] - India,
Garo [grt] - India,
Gikuyu [kik] - Kenya,
Goan Konkani [gom] - India,
Godwari [gdx] - India,
Gogo [gog] - Tanzania,
Gondi [gon] - India,
Gorontalo [gor] - Indonesia (Sulawesi),
Guarani [grn] - Paraguay, Bolivia,
Hadiyya [hdy] - Ethiopia,
Haitian [hat] - Haiti,
Halh Mongolian [khk] - Mongolia,
Haryanvi [bgc] - India,
Hassaniyya [mey] - Mauritania,
Hausa [hau] - Nigeria,
Haya [hay] - Tanzania,
Hazaragi [haz] - Afghanistan,
Hiligaynon [hil] - Philippines,
Hmong Daw [mww] - China,
Hmong [hmn] - China,
Ho [hoc] - India,
Hunsrik [hrx] - Brazil,
Ibibio [ibb] - Nigeria,
Igbo [ibo] - Nigeria,
Ilocano [ilo] - Philippines,
Izon [ijc] - Nigeria,
Jambi Malay [jax] - Indonesia,
Javanese [jav] - Indonesia (Java and Bali),
Jula [dyu] - Burkina Faso,
Kabardian [kbd] - Russian Federation,
Kabuverdianu [kea] - Cape Verde Islands,
Kabyle [kab] - Algeria,
Kalenjin [kln] - Kenya,
Kamba [kam] - Kenya,
Kanauji [bjj] - India,
Kangri [xnr] - India,
Kannada [kan] - India,
Kanuri [kau] - Nigeria,
Kashkay [qxq] - Iran,
Kashmiri [kas] - India,
Khams Tibetan [khg] - China,
Kimbundu [kmb] - Angola,
Kimîîru [mer] - Kenya,
Kipsigis [sgc] - Kenya,
Kituba [ktu] - Dem. Rep. of Congo,
Kituba [mkw] - Congo,
Kongo [kon] - Dem. Rep. of Congo,
Konkani [knn] - India,
Koongo [kng] - Dem. Rep. of Congo,
Kpelle [kpe] - Guinea,
Kumaoni [kfy] - India,
Kurdish [kur] - Kurdistan, Iraq, Turkey,
Kurux [kru] - India,
Kyrgyz [kir] - Kyrgyzstan,
Lahnda [lah] - Pakistan,
Laki [lki] - Iran,
Lambadi [lmn] - India,
Lango [laj] - Uganda,
Lao [lao] - Laos,
Limburgish [lim] - Netherlands,
Lingala [lin] - Dem. Rep. of Congo,
Lomwe [ngl] - Mozambique,
Luba-Kasai [lua] - Dem. Rep. of Congo,
Luba-Katanga [lub] - Dem. Rep. of Congo,
Lubukusu [bxk] - Kenya,
Lugbara [lgg] - Uganda,
Maasai [mas] - Kenya,
Maasina Fulfulde [ffm] - Mali,
Madura [mad] - Indonesia (Java and Bali),
Magahi [mag] - India,
Maguindanao [mdh] - Philippines,
Mahasu Pahari [bfz] - India,
Maithili [mai] - India,
Makasar [mak] - Indonesia (Sulawesi),
Makhuwa [vmw] - Mozambique,
Makhuwa-Meetto [mgh] - Mozambique,
Makonde [kde] - Tanzania,
Malagasy [mlg] - Madagascar,
Malay [msa] - Thailand, Malaysia,
Malay [zlm] - Malaysia (Peninsular),
Malayalam [mal] - India,
Mandingo [man] - Senegal,
Mandinka [mnk] - Senegal,
Manyika [mxc] - Zimbabwe,
Marwari [mwr] - India,
Masaaba [myx] - Uganda,
Mazanderani [mzn] - Iran,
Meitei [mni] - India,
Mende [men] - Sierra Leone,

Min Dong Chinese [cdo] - China,
Min Nan Chinese [nan] - China,
Mina [myi] - India,
Minangkabau [min] - Indonesia (Sumatra),
Mundari [unr] - India,
Muong [mtq] - Viet Nam,
Musi [mui] - Indonesia (Sumatra),
Mòoré [mos] - Burkina Faso,
Ndau [ndc] - Zimbabwe,
Ndebele [nde] - Zimbabwe,
Ndonga [ndo] - Namibia,
Ngbaka [nga] - Dem. Rep. of Congo,
Nigerian Fulfulde [fuv] - Nigeria,
Nigerian Pidgin [pcm] - Nigeria,
Nimadi [noe] - India,
Northern Hindko [hno] - Pakistan,
Northern Khmer [kxm] - Thailand,
Northern Luri [lrc] - Iran,
Northern Qiandong Miao [hea] - China,
Northern Sotho [nso] - South Africa,
Nuosu [iii] - China,
Nyakyusa-Ngonde [nyy] - Tanzania,
Nyanja [nya] - Malawi,
Nyankore [nyn] - Uganda,
Occitan [oci] - France,
Oluluyia [luy] - Kenya,
Oriya [ori] - India,
Oromo [orm] - Ethiopia,
Pahari-Potwari [phr] - Pakistan,
Pampangan [pam] - Philippines,
Pangasinan [pag] - Philippines,
Pontic [pnt] - Greece,
Pulaar [fuc] - Senegal,
Pular [fuf] - Guinea,
Pwo Eastern Karen [kjp] - Myanmar,
Quechua [que] - Bolivia, Peru,
Quiché [quc] - Guatemala,
Rajasthani [raj] - India,
Rangpuri [rkt] - Bangladesh,
Rohingya [rhg] - Myanmar,
Romany [rom] - Romania,
Rundi [run] - Burundi,
Rwanda [kin] - Rwanda,
S'gaw Karen [ksw] - Myanmar,
Sadri [sck] - India,
Santali [sat] - India,
Sasak [sas] - Indonesia (Nusa Tenggara),
Sena [seh] - Mozambique,
Seraiki [skr] - Pakistan,
Serer-Sine [srr] - Senegal,
Shan [shn] - Myanmar,
Shekhawati [swv] - India,
Shona [sna] - Zimbabwe,
Sidamo [sid] - Ethiopia,
Sindhi [snd] - Pakistan,
Sinhala sin [Sri] - Lanka,
Soga [xog] - Uganda,
Somali [som] - Somalia,
Songe [sop] - Dem. Rep. of Congo,
Soninke [snk] - Mali,
Southern Balochi [bcc] - Pakistan,
Southern Dong [kmc] - China,
Southern Kurdish [sdh] - Iran,
Southern Ndebele [nbl] - South Africa,
Southern Sotho [sot] - South Africa, Lesotho,
Sukuma [suk] - Tanzania,
Sunda [sun] - Indonesia (Java and Bali),
Surgujia [sgj] - India,
Surjapuri [sjp] - India,
Susu [sus] - Guinea,
Swahili [swa] - Tanzania,
Swati [ssw] - South Africa, Swaziland,
Sylheti [syl] - Bangladesh,
Tachawit [shy] - Algeria,
Tachelhit [shi] - Morocco,
Tagalog [tgl] - Philippines,
Tajiki [tgk] - Tajikistan,
Tamashek [tmh] - Niger,
Tarifit [rif] - Morocco,
Tausug [tsg] - Philippines,
Teso [teo] - Uganda,
Thai [tha] - Thailand,
Themne [tem] - Sierra Leone,
Tibetan [bod] - China,
Tigrigna [tir] - Ethiopia, Eritrea,
Tigré [tig] - Eritrea,
Tiv [tiv] - Nigeria,
Tonga [toi] - Zambia, Zimbabwe,
Tsonga [tso] - South Africa,
Tswa [tsc] - Mozambique,
Tswana [tsn] - South Africa, Botswana,
Tulu [tcy] - India,
Tumbuka [tum] - Malawi,
Turkmen [tuk] - Turkmenistan,
Tày [tyz] - Viet Nam,
Umbundu [umb] - Angola,
Uyghur [uig] - China,
Uzbek [uzb] - Uzbekistan,
Varhadi-Nagpuri [vah] - India,
Vasavi [vas] - India,
Venda [ven] - South Africa,
Vlaams [vls] - Belgium,
Waray-Waray [war] - Philippines,
Western Balochi [bgn] - Pakistan,
Western Panjabi [pnb] - Pakistan,
Wolaytta [wal] - Ethiopia,
Wolof [wol] - Senegal,
Xhosa [xho] - South Africa,
Yao [yao] - Malawi,
Yilumbu [lup] - Gabon,
Yombe [yom] - Dem. Rep. of Congo,
Yoruba [yor] - Nigeria,
Zande [zne] - Dem. Rep. of Congo,
Zarma [dje] - Niger,
Zaza [zza] - Turkey,
Zhuang [zha] - China,
Zulu [zul] - South Africa