A Statistical Typology of (Textual) Language in Finer Granularity

Anonymous ACL submission

Abstract

We propose a character-level perspective for a new understanding and visualization of language, in its textual representation in computing, using relative line length and character vocabulary size from parallel corpora as parameters. We discover an emergent pattern with a natural, continuous order to languages. We highlight some of the outlier languages and discuss the opportunities and challenges in line for character- and byte-level development in language technology.

1 Introduction

In recent years, advances in neural machine learning (ML) have shown great success with representations that are more fine-grained than *word*. Results from character and byte language models (LMs) have become on par with those from word models, mitigating crosslinguistic performance disparity that can negatively impact languages morphologically rich and poor (Wan, 2021). Mielke et al. (2019) examined character and Byte-Pair-Encoding (Sennrich et al., 2016) LMs and Wan (2021) character-, byte-, and word-level conditional LMs and both concluded basic data statistics in vocabulary size and line length to be correlated with performance. In this non-experimental data analysis and visualization paper, we explore the understudied subword space in language technology by illustrating:

- the character profile of the world's languages[1],
- our discovery of a novel, continuous quasi-sigmoidal natural order,
- some opportunities and challenges for character- and byte-level development in language technology.

2 Language profile in characters

We represent each language in a parallel corpus using its **mean line length in characters** and its **total vocabulary size ($|V|$) in characters** for the corpus. As standard in computing, *line*[2] is defined as a sequence of characters delimited by a carriage return or line feed. A *character* is a "basic unit of encoding for the Unicode character encoding"[3]. We graphed these two simple statistics, mean line length in characters on the x-axis and character $|V|$ on the y-axis, for each of the following three multi-way parallel corpora ("multitexts")[4]:

- the Universal Declaration of Human Rights (UDHR) from the *UDHR in Unicode* project[5]: we used full texts of all 460 languages (Fig. 1);
- the Bible data from Christodoulopoulos and Steedman (2015)[6] in 108 unique languages (C&S Bible) (Fig. 2); and
- the Bible data from Mayer and Cysouw (2014) in 1168 varieties (M&C Bible) (Fig. 3).

Note: the text (language labels) in the figures are not meant to be readable — what we wanted to show is the contrast between a majority group and the outliers.

[1] as accessible through available parallel corpora

[2] "Line" is used here since the notion of "sentence" is not well-defined for all languages, esp. the linguae/scriptiones continuae (languages without explicit sentence markers, such as Thai). It is beyond our scope here to define best practices in crosslinguistic sentential segmentation and alignment, but conventionally a sentence is at least a line, but longer ones can span several lines. For the purpose of this paper, we assume that by the time parallel data is inputted into a machine, any relevant "sentence"-like units would have been converted to lines, demarcated by a newline character or the like.

[3] from https://unicode.org/glossary/

[4] A complete list of all language names/codes is provided in Appendix A.

[5] https://www.unicode.org/udhr/aggregates.html

[6] https://github.com/christos-c/bible-corpus

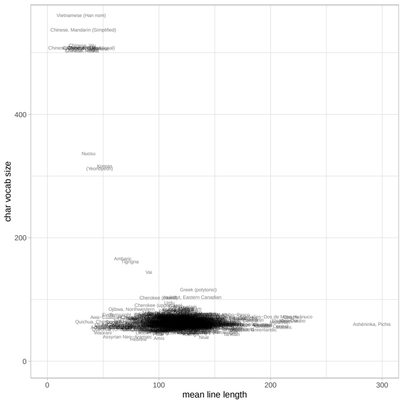

Figure 1: UDHR

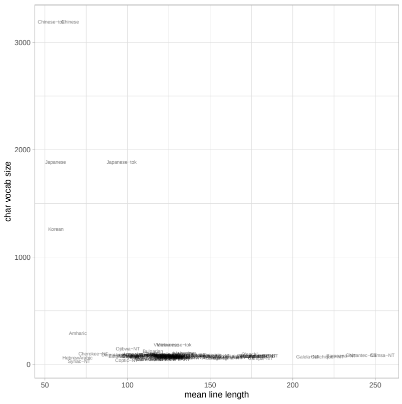

Figure 2: C&S Bible

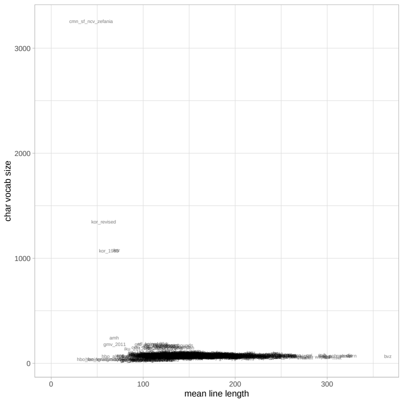

Figure 3: M&C Bible

2.1 The outliers in character $|V|$ and length

A common pattern emerges from Figures 1, 2, and 3. The majority of these languages cluster into a bulk in the center, there are some outliers that are high in character $|V|$ but low in line length and some that are low in line length but high in $|V|$. But there are no languages that are high in both.

Logographic, as opposed to (primarily) phonetic alphabetic, languages[7] such as Vietnamese (Han Nom), varieties of Chinese, Nuosu, and Korean (as well as Japanese from the C&S Bible data) are all languages that are high in vocabulary but low in average line length. They are overlooked not because they are low-resource languages or less studied (although Nuosu would be one), but because their distinctiveness can often be neutralized/camouflaged with a phonetic script: whereas Vietnamese in traditional Han Nom script is an outlier, Vietnamese in the Latin script is not.

For the UDHR dataset, there are 10 extreme outliers with $|V|$ of over 500: Vietnamese (Han Nom), all ZH languages, of which there are 8 in this dataset — Mandarin (Simplified), Wu, Gan, Mandarin (Traditional), Min Nan, Yue, Jinyu, Hakka, and Japanese. The next ones are Nuosu, Korean, and Yeonbyeon with $|V|$ in the 300s. 12 of these 13 also have the shortest mean length (about 30+) of all the languages in this dataset. The outlier with the highest mean line length of 291[8] is Ashéninka (Pichis).

For the C&S Bible, the outliers in $|V|$ are Chinese and ZH tokenized (ca. 3000), Japanese and JA-tok (almost 2000), and Korean (1259). The next language in rank, Amharic, shows a significant drop to 289. The outliers in length are Camsa-NT (New Testament), Chinantec-NT, Barasana-NT, Cakchiquel-NT, and Galela-NT with mean length between 209 and 254. All other languages are below 200. 4 of the 5 outliers with highest $|V|$ are also at the bottom of the list when length is ranked in descending order (with 24-29 characters).

For the M&C Bible, Chinese and Korean are again on top of the list with $|V|$ about 3k and 1k higher than the rest and again on the bottom of

[7] or "languages with logographic scripts" — although the shorter formulation can also be used as a short hand, the difference in the two wordings conveys a subtle difference in how a human writer/speaker relates the concept of logography to language. This would make for an interesting discussion in its own right. However, neither orthography nor writing systems are a subject of relevance here. We are concerned with language represented as text in computing.

[8] Figures rounded to whole number.

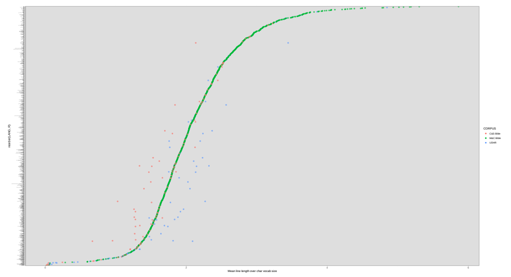

Figure 4: All 3 datasets combined with the ratio of mean line length and character $|V|$ on the x-axis. On the y-axis are languages ranked by this ratio in descending order. Languages present in multiple datasets get sorted according to the mean of the ratios across datasets — this surfaces as dots "off" the main curve. (Figure is to show general tendency. Language labels are not meant to be legible here, but are listed in App. A.)

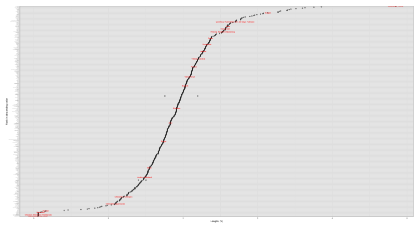

Figure 5: Languages of UDHR ranked in a continuum with the ratio of length and character size.

the list when length is sorted in descending order. Highest mean length is about 366 for bvz (Bauzi) and values decrease rather consistently thereafter with little obvious jump across 1168 entries.

These data seem to suggest that languages fall into two broad clusters of logographic and phonetic languages/scripts. But considering how the first 2 of these datasets are the only existing, completely *multiway* parallel ones (not just bitexts) which include at least 2 CJKV languages, there is room for this hypothesis to be further examined. We hope and call for the community's attention in the continual creation of these resources.

3 A natural order

Figure 4 illustrates the **ratio** of mean line length and character $|V|$ ($\frac{\text{length}}{|V|}$) on the x-axis, and on the y-axis, 1693 language varieties from the 3 datasets combined ranked in descending order of such ratio. We see a quasi-sigmoidal curve indicating that languages can be viewed as ordered in a continuum instead of being classified into disparate rigid types. The shape remains even when data from only one corpus were plotted, as in Fig. 5, suggesting that this emergent phenomenon is robust and is one that transcends data-specific details. We see that there is a structure to language also on the character level. (Note: all our figures illustrate patterns derived from raw data without explicit segmentation, i.e. whitespaces are accounted for in languages that do have them; languages that don't have them naturally are represented so in UDHR, some in both word-tokenized and untokenized form across the Bible data. We also did not use any dimensionality reduction.)

Although there may have been a historical tendency in linguistic typology to classify (whole) languages, as opposed to specific linguistic phenomena, into three or four types, e.g. morphological types such as isolating, fusional, introflexive, and agglutinative, "the majority (perhaps all) of the world's languages do not correspond exactly to one or other of these types" (see Comrie (1981), among others). A concept of continuity and flexibility in linguistic analyses has stood long in the implicit understanding of many linguists. While our discovery of a continuous profile of languages expressed as the relationship between rank and the ratio of length and character $|V|$ (Figure 4) is novel, similar to the discovery of patterns illustrating natural laws such as Zipf's (Zipf, 1935), our effort can also be seen as mirroring this implicit understanding of symbolic continuity in the computational context. With each parallel corpus, the rank of language varieties can vary and be used for relative comparison with others. And this method of comparison can also be extended to contrast sets where variation in genre and style within one variety can be studied. The ratio can be used as a convenient index to gauge statistical profiles, enabling a more flexible comparison of the world's language varieties as a form of practice in comparative language science as well as a more succinct assessment of language data profiles in general.

4 Opportunities and challenges for more fine-grained development

Complement character encoding design with languages' statistical profiles To help facilitate algorithmic fairness in crosslingual/multilingual work, languages that tend to have longer line lengths can benefit from compression and those shorter, from decomposition. Equitable measures for more complementarity between characters and bytes through better character encoding design can be taken to compensate the differences in distribution (see Fig. 6), serving non-neural appli-

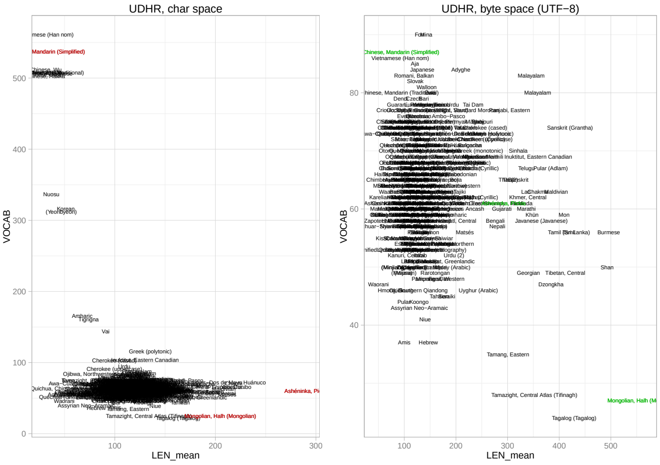

Figure 6: UDHR in character space (left) and byte space (UTF-8) (right), with mean line length on the x-axis and $|V|$ on the y-axis in characters and bytes, respectively. The left subfigure is the same as Fig. 1 but with three outlier languages highlighted in red. To create fairer representation for computation, we could bring the outliers closer to the cluster. Chinese (Mandarin, simplified) is an outlier in both spaces, highlighted in top left in red in char space and green in byte space. Ashéninka (Pichis) is long with low character $|V|$ (red on left) and we see that it is rather "well-integrated" in byte space. Mongolian (Halh) could still use a lower byte count.

cations and situations/locations where computational/technological resources are scarce.

Parallel data with CJKV The UDHR dataset is by far the most diverse parallel dataset encompassing the broadest range in script varieties and is, to the best of our knowledge, the only dataset that contains fully parallel content for all native CJKV scripts aligned with other major languages. It is, however, meager in size and has only a few hundred lines for each language variety. Even though our study already considers more languages than almost all multilingual papers nowadays, it is still just a small fraction considering there are around 7k recognized language varieties in the world. The coverage of publicly available datasets is far from being inclusive and diverse on a global scale.

Alignment of byte units and subcharacter information for logographic languages Alignment is essential for the automatic compilation of lexicographic resources. For logographic languages, a character can be decomposed into subcharacter-level semantic and phonetic elements (see e.g. Zhang and Komachi (2018)). Alignment algorithms are plenty, but the elementary units on which these algorithms are to be applied still need work. Finer-grained alignment between sub-character components in logographic languages and those in other logographic languages or sub-"word" components (i.e. character or sub-character strings, depending on the script) in non-logographic languages is still an understudied research area to which more intelligent decomposition of sub-character units can be a qualitative contribution.

Language documentation and digitization "Alphabetization" has happened as an artefact of cultural imperialism and affected languages such as the Egyptian Hieroglypics, Mi'kmaq, Chinese, and Vietnamese. We can help communities document, digitize, and preserve indigenous writing. The degree of threat and endangerment has not been monitored and calibrated for orthography/scripts, only for "languages". Though downstream performance disparity of logographic languages can often be mitigated with transliteration, it can be a cultural matter meaningful to the language communities concerned if/when we do take their native (often less ASCII-like) representation into account.

5 Conclusion

In this paper, we showed some novel perspectives in conceptualizing languages on the character level using basic statistics, inspired by experimental findings in ML. The deployment of more scalable ML methods can often lead to more languages being included in NLP research (Joshi et al., 2020), and this paper is a first attempt to reconcile some of the many opportunities available for qualitative development in this direction.

References

Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Bernard Comrie. 1981. *Language Universals and Linguistic Typology*. Blackwell, Oxford.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3158–3163, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ada Wan. 2021. Representation and bias in multilingual NLP: Insights from controlled experiments on conditional language modeling.

Longtu Zhang and Mamoru Komachi. 2018. Neural machine translation of logographic language using sub-character level information. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 17–25, Belgium, Brussels. Association for Computational Linguistics.

George Kingsley Zipf. 1935. *The Psychobiology of Language*. Houghton-Mifflin, New York, NY, USA.

A Complete list of languages used

Language names and language codes as used by the datasets analyzed. In total, there are $1,693$ unique label values. They are ordered by mean line length in characters divided by character vocabulary size, in descending order, as ranked from highest to lowest in Figure 4.

mnx, bvz, mwf, hix, aoi, mav, akh, Ashéninka_Pichis, wmt, wbp, ipi, auc, xav, agm, opm, poh_western, djr, ubu_nopenge, tlf, nab, sua, apz, rwo_karo, big, yaq, sab, kyg, rwo_rawa, ino, ura, pmf_taa, aso, gdn, bef, gvf, Matsés, cbt, mux, amm, Tahitian, omw, iws, hin, grc_textusreceptusVAR2, ese, wim, mej, guh, mri, azz, stp, hnn, nak, ssx, cui, grc_unaccented, Inuktitut_Greenlandic, cak_yepocapa, ubu_kala, tos, Niue, waj, hch, mpt, kyz, chr, ahk, mee, gnn, Nahuatl_Central, bis, bhl, aon, sny, abt_maprik, zam, Arabela, kmu, mtj, qup, aby, Rarotongan, Toba, abt_wosera, yml, arl, toc, pio, spl, nvm, tee, rop, Maldivian, snn_2009, Yanesha, nii, mta, cta, cok, kwf, cni, Shipibo-Conibo, Caquinte, ape, crn, etr, acd, viv, mal, ncu, rkb, Campa-NT, imo, trc, wiu, dtp_kadazan, cak_western, apn, roo, mks, nuy, bvr, Cakchiquel-NT, grc_textusreceptusVAR1, cjv, nlc, grc_byzantine, Tongan, yli, byx, ame, wrs, ban, bbb, muh, grc_tischendorf, for, ncl, kup, mcb, gnd, iry, yby, Chachi, Shan, sgz, wnc, lhu, lac, mps, fai, cjo, cak_central1980, bku, agg, sll, Chin_Haka, faa, tzj_western, txu, cav, grc_wescotthortVAR2, grc_wescotthortVAR1, grc_combined2005, twu, Pampangan, bon, gdr, gbi, ycn, xbi_wampukuamp, dnj_eastern, Samoan, cot, tte, mbb, kto, azg, urk, tpi, Achuar-Shiwiar, hus_sanluispotosi, sim, snp_lambau, tac, ctp, xbi_yanimoi, sbl, apu, nsn, byr, aia, hui, gfk, Koongo, amp, way, Galela-NT, dah, Pidgin_Nigerian, Camsa-NT, pot, snp_komongu, tpz, mva, gor, yva, kze, lif, yss_yawu, tim, Chinantec-NT, cco, urb, snn, Uyghur (Latin), mto, xsi, lwo, kzf, kbm, kac, qvs, Konjo, ptu, soq, aak, boj, aom, uvl, mzl_2005, Barasana-NT, xla, min, kpx, srm, Totonac_Papantla, aau, Maori, hvn, Uyghur (Arabic), kwj, mox, ppk, yuw, Micmac, hmo, Amahuaca, kjs, kno, Aguaruna, gah, ong, gvn, ata, aoj_mufian, yuj, plg, kxm, rai, npy, smo, dnj_western, mhx, lid, ppo, tuc_ont2011, Swati, pad, cak_eastern, pse, naf, Quechua_Northern Conchucos Ancash, Asháninka, Amis, mak, tby, zlm_leydekker, sps, Quechua_Huamalíes-Dos de Mayo Huánuco, med, kbh, stn, bcw, pis, bjr, sco, aoj_filifita, tzj_eastern, ifb, mop, Georgian, tuf, trq, quc_joyabaj, ntp, awb, uig, cso, mlp, pah, udu, ium_roman, ind_terjemahanlama, alq, amx, kac_bsm, Hausa (Nigeria), gil, tuc_2011, Hausa (Niger), mmx, pne, Ilocano, ium_ancientroman, mad, nas, Pintupi-Luritja, bwu, mbt, kgk, caa, Burmese, Tamil (Sri Lanka), Tamil, Bora, gym, Kaqchikel_Central, auy, jac, yrb, bmh, mnh, maq, Bicolano_Central, mnb, toj, guo, noa, ifu, Bali, bmu, Marshallese, kyc, mif, msy, Ibibio, tgl_1905, Chin_Falam, Chin_Tedim, djk, nch, Hmong Njua, zpv, ian, mek, Hiligaynon, cak_southern, Navajo, kue, ghs, Hmong_Southern Qiandong, Kituba, dob, jvn, ksr, Gaelic-PART, tzo_huixtan, Uma-NT, mvp, amr, Javanese (Latin), mpm, Potawatomi-PART, kmh, Akawaio-NT, Sotho_Northern, ctu_tili, lem, yad, cnl, cof, kms, cbc, Aukan-NT, hlt, som, ium_lao, Bugis, maj_2011, lia, apy, snc, bbr, bug, Tagalog, qvi, nwi_2012, kew, cak_santamaria, agn, cbu, cbr, pir, wer, ntr, bjp, ind_terjemahanbaru, mmo, top, mam_todossantos, hla, guc, acr_cubulcu, obo_2011, mhl, nss, iou, myw, nhw, mya_common, cap_1978, cop_bohairic, avu, zia, jic, avt, mlh, moc_2005, bps, Malagasy_Plateau, yss_mayo, mzz, jav_1981, tbc, mfi, lsi_2009, xrb, haw, Tojolabal, Romansch (Surmiran), mwv, Coptic-NT, kpr, quc_westcentral, cux, hus_veracruz, Lobi, maw, qvc, kpw, yon, arn, kud, wnu, mbf, khz, car, mio, ind_kabarbaik, Edo, yle, zav, coe, ake, ium, cop_sahidic, wap, ixl, mau, pib, ifk, dgz, ign_2004, too, tlb, agd, Tsonga (Zimbabwe), Romansch (Sursilvan), smk, fil_1905, Aceh, tna, Somali, Bislama, lcm, tbo, tav, tam, inb, tcs, itv, gaw, nij, enb, kvn, kqw, Ladino, nou, Baoulé, cle, Comorian_Ngazidja, Comorian_Maore, bnp, Syriac-NT, Tswana, cbv, blz, K'iche'-NT-SIL, mjc, Seselwa Creole French, Pipil, hig, arz, pls, ncj, K'iche'_Central, afr_boodskap, Cashibo-Cacataibo, Malay (Latin), mna, ceb_bugna2009, btd, szb, fij_1974, Lingala (tones), knj, mir, nhe, Bushi, maz, hub_2010, mie, chd, khm, Minangkabau, tbl, bsn, Galician, sxn_sangir, cut, nia, myy, ter, Sunda, Nahuatl-NT, Pohnpeian, krj, Mozarabic, Hmong_Northern Qiandong, dts, pwg, qul_2006, bmr, nif, Romansch (Grischun), zpc, chf, quh_chumacharazani, Shuar, gog, acn, Gaelic_Irish, Romansch (Sutsilvan), ell_modern2009, Haitian Creole French (Popular), zlm_todaysmalay, otn, usa, cub, Romansch, cya, amn, con, mil, gvc, agu, Shuar-NT, sda, kaq, Jakalteko-NT, amu, ssd, Mbundu (009), Pular, Nyemba, dwr, yut, pbi, zac, mbl, tlh_klingon, Waray-Waray, neb, hub_1975, nld, bkq, cao, Palauan, Khasi, mcq, wbm, fin_1933, usp, qvz, Amuzgo-NT, huu, kmo, ngu, mhy, hbo_leningradunicode, nhy, Tzeltal_Oxchuc, spy, tcc, Drung, ceb_pinadayag, cax_1980, hto, Sardinian_Logudorese, mxp, Cebuano, tzo_chenalho, Tachelhit-NT, epo, crq, mfz, are, Yanomamö, Achuar-NT, ace, mog, emi, Estonian, jav, Lingala, Quechua_South Bolivian, mbj_2011, msa_1996, Uspanteco-NT, xon_likpakpaaln, wob, Tibetan_Central, Indonesian, (Bizisa), jiv, csk, dww, Yiddish_Eastern, Nyanja (Chechewa), Mískito, zpz, Slovenian, ctd, tzo_zinacantan, cbs, bto, acu, Crioulo_Upper Guinea, Romansch (Puter), acr_reformed, mpx, hag, Interlingua, acr_traditional, Tzotzil (Chamula), aey, eng_amplified, sxn_siau, mzk, mnk, mwc, Páez, Malay (Arabic), hbo_westminster, Sotho_Southern, mco, Romansch (Vallader), bts, cnh, mxt, nso_2000, xon_likoonli, msm, des, sgb, bzj, cko, Madura, Chin_Matu, sue, Basque, ceb_bugna, sot_southern1989, leu, Romanian (1953), kyq, sun, Wolaytta-NT, mih, pao, poh_eastern, hwc_2000, taj, yua, ljp, tzo_chamula, Maya_Yucatán, gde, Oroqen, nyf, Chuukese, ded, Finnish_Kven, dug, Quechua_Margos-Yarowilca-Lauricocha, zar, Quechua_North Junín, bex, cfm, Luvale, mxq, kqf, Quechua_Ambo-Pasco, Manx, sxb, iba, Romanian (1993), Twi (Asante), mam_central, Romanian (2006), Amarakaeri, gum, Papiamentu, Picard, zat, nhg, quf, deu_greber, Candoshi-Shapra, Záparo, hns, ann, Latin, Luba-Kasai, heh, xsm, suk, eng_literal, kma, pam, ptp, Yoruba, Fijian, English, mlg_1865, Chamorro, yal, mfh, eng_diaglot, kng, Gaelic_Scottish, ind_suciinjil, mib, ify, eng_kingjames, xtd, ign_1980, poe, Yao, lif_2009, Mam-NT, btx, mya, mig, Venetian, Purepecha, Saxon_Low, zty, zos, fin_1766, nop, mxb, Hawaiian, zpl, Nzema, plt_interconfessional, cak_central2003, Latin (1), fij, Uzbek_Northern (Latin), kek_1988, cac_sansebastian, Mongolian_Halh (Cyrillic), kgp, aui, kpg, (Minjiang_spoken), quy, trn, tca, quh_1976, cpa, zpq, sil, ron_cornilescu, pag, Mbundu, Bamun, ind_easy2005, Chayahuita, lef, Dutch, sig, bmk, ita_diodati, spy_2011, wsk, Albanian_Tosk, Wayuu, (Maiunan), tzo_chamula2000, Huastec (Sierra de Otontepec), cuc, gwi, eng_darby, mur, Saami_North, bzh, Quechua_Ayacucho, Finnish, rmy, ixl_2001, kkj, swh, Greek, Corsican, agr, kat, kri, Aromanian, sus, awi, Dzongkha, deu_elberfelder1905, rro, (Minjiang_written), bcl, Portuguese (Portugal), (Mijisa), sot_southern1961, pkb, cpu, cnw, thk, Secoya, Mon, Mizo, ubr, heb_2009, crx, eus_batua, Frisian_Western, ote, afr_viralmal, nso_neworthography, Farsi_Western, eng_montgomery, Dagaare_Southern, Aguaruna-NT, ood, awx, Javanese (Javanese), Dangme, kpf, toh, tpm, hak, otm, Venda, tku, Welsh, knk, ttc, tpt, Chokwe, men, Mòoré, dyo, Norwegian_Bokmål, ndo_2008, Ndonga, gui, poi, Myanmar, hot, wal, pbc, Hani, nfr, cac_ixtatan, Ese Ejja, prf_2012, xho, qvn, Greek (monotonic), cax_2002, Portuguese (Brazil), maa_sanjeronimo, sna, cuk, Lukpa-NT, Nganasan, Bemba, deu_konkordant, tsn_1908, Urarina, cwt, Pijin, Kirghiz, Esperanto, las, dop, tuo, Yapese, aln, hil, tpa, tgp, myu, zpm, Umbundu, jam_2012, Kanuri_Central, irk, Quechua_Huaylas Ancash, Occitan (Francoprovençal_Valais), qug_chimborazo2010, dip, war, Asturian, Nomatsiguenga, Oromo_Borana-Arsi-Guji, prs, sur, taq, Rundi, zsr, qxr, Scots, Abkhaz, Occitan (Francoprovençal_Savoie), Huitoto_Murui, bao, Themne, Romagnolo, Norwegian_Nynorsk, Tiv, nya, Occitan (Francoprovençal_Fribourg), Xhosa, vmy, nca, Dari, swp, Koongo (Angola), Altai_Southern, tsw, ita_riveduta, Khakas, buk, chz, huv, Veps, kgf, bav, lus_1959, mqb, qxn, boa, Malagasy, Occitan (Languedocien), dig, csy, glv, sri, eng_majority, tzh_oxchuc, Italian, aai, zao, Mam_Northern, Tsonga (Mozambique), teo, pbb, pab, shi, kdl_2010, quh_1993, qug_chimborazo, sja, nde_2012, spa_tla, Moba, vun, hbo_leningradconsonants, hbo_leningradunicodeconsonants, mfq, kus, okv, Afrikaans, cha_2003, mti, kki, tzo_1997, mam_southern, kqp, fra_louissegond, Tonga, Hindustani_Sarnami, kao, Tem, nhi, Dinka-NT, Sorbian_Upper, qxo, nob_1988, Tok Pisin, deu_interlinear, Kabardian, Manx-PART, Okiek, deu_luther1912, sqi, Kabuverdianu, Occitan, Zapotec_Güilá, ldi, Zulu, ndz, lat_vulgataclementina, Hungarian, Zhuang_Yongbei, German_Standard (1996), French, eus_Hautin1571, Quechua_Cajamarca, Romanian, German_Standard (1901), zpi, ekk_1968, Orok, kjb, eus_navarrolabourdin, quc_1995, Swahili-NT, cym_morgan1804, syc, Ganda, bpr, hat_1999, afr_1953, quz, gux, wuv, Turkmen (Cyrillic), Nyanja (Chinyanja), cce, zpu, K'iche'-NT, hye_latin, fas_1995, Turkmen (Latin), bib, eng_etheridge, zpo, Romani_Balkan (1), cme, Yagua, gur_frafra, Q'eqchi', Paite, Spanish, kek_2005, myk, bgr, rim, srn, czt, zad, deu_erben, ilo, ita_nuovadiodati1991, esi, deu_luther1545, njm_1970, Assyrian Neo-Aramaic, mit, eng_newreaders, deu_luther1545letztehand, daa, gyr_1985, Yakut, Bari, Zarma, Ticuna, jmc, Huastec (Veracruz), Tamazight_Central Atlas, fra_kingjames, lat_novavulgata, nim, Ga, Swahili, cym_revised2004, Rwanda, deu_hoffnung, hat_1985, wed_topura, Swedish, eng_easytoread, Albanian, Basque-NT, gmv, Maltese, Garifuna, Éwé, eng_worldwide, zab, dik_dinka, Ligurian, qul_1985, Twi (Akuapem), esk, eng_lexham, kal, dje, sum, Igbo, Krio, khm_2011, dan_frederik, Luxembourgeois, zas, maa_sanantonio, Tuva, Shona, ceb_popular, Guarayu, blh, mfe_2009, bjv, bru, afr_1983, bre, Mende, ade, deu_elberfelder1871, qvh, Quechua (Unified Quichua_old Hispanic orthography), nnw, ava, nor_student, Friulian, nhu, Lao, Huastec (San Luís Potosí), jbu, qwh, Haitian Creole French (Kreyol), yam, fra_geneve1669, Pashto_Northern, gur_ninkare, qxh, mda, swh_union, mzw, Tajiki, English-WEB, Occitan (Francoprovençal_Vaud), chq, dik, gnw, qub_2009, fra_darby, dad, pan, Nyankore, fas_2007, Danish, Ladin, srq, ita_vita1997, sme, Bosnian (Latin), spa_reinavalera1960, eng_newcentury, urd, kan, Limba_West-Central, Tetun, swe_nyalevande, Kurdish_Northern, Kurdish_Central, nob, vag, deu_freebible, tsn_1970, ekk_1997, ayr_1997, kwi, bxr, cnt, pes, Ditammari, eng_clontz, dan_hverdagsdansk, Thai-tok, zul, deu_tafelbibel, Catalan-Valencian-Balear, qvm, nld_2004, Lozi, spa_dioshablahoy, otq, Sukuma, fra_semeur, myv, tso, anv, Serbian (Latin), zaw, zai, Dagbani, fil_2005, deu_reinhardt1910, Macedonian, fin_1992, plt_romancatholic, Belarusan, gyr_2002, Sharanahua, swh_habarinjema, mam_northern, tgl_1996, gyr_urubicha, kum, ctu_tumbala, izr, eng_goodnews, Azerbaijani_North (Cyrillic), Breton, Afar, Azerbaijani_North (Latin), bwq, Uzbek_Northern (Cyrillic), Tetun Dili, tzh_bachajon, Fante, eng_new2007, fra_ostervald1867, deu_textbibel, sqi_interconfessional, sqi_2007, acf, cym, nld_1951, eng_newsimplified, hau, Quechua_Arequipa-La Unión, ita_2009, Hebrew, Gagauz, tik, kqc, old, miq, mcu, Polish, eng_newliving, eng_books, Bosnian (Cyrillic), pbl_1994, npl, wed_wedau, afr_lewende, Chakma, Walloon, Latvian, urd_arabic, deu_schlachter, pol, deu_albrecht, nld_2007, Panjabi_Western, tem, kaz, deu_zuercher, nob_1930, Baatonum, Fulfulde_Nigerian (2), fra_davidmartin, Kabyle-NT, por_linguagemdehoje, gof_2011, Mapudungun, kin_bird_youversion, kin_2004, pua_2011, ita_nuovariveduta, che, urd_devanagari, fas_newmillennium2011, eng_common, hun_2012, Fulfulde_Nigerian, Nanai, fra_courant1997, mos_protestant, kub, kin_2012, Creole, Turkish, ron_2006, miz, bth, ayr_2011, Sango, Occitan (Auvergnat), bul_veren, Serbian (Cyrillic), tfr, myb_1980, cha_2010, Kasem, wol, Makhuwa, Urdu (2), Russian, spa_blph_youversion, ory, dyu, Colorado, zom, Romani-NT, oss, cjp, German, Ukrainian, ben_holybible, tcw, tur, Crimean Tatar, Jula, ben_mussolmani, swe_folk1998, hif, fra_bonnet, hun_karoli, hay, tue, mzm, cha_1908, Gujarati-NT, spa_rvr95_youversion, fal, spa_dhhe, ztq, Seraiki, tur_2009, bam, eng_livingoracles, Lunda, emp, Croatian, eng_godsword, Farsi, eng_contemporary, pol_gdansk, Nepali, bba, cap_2004, deu_pattloch, Talysh, tab, ceb_godsword, Bamanankan, nus, Sanskrit (Grantha), Icelandic, Kissi_Northern, Mazatec_Ixcatlán, Kannada, Wolof, dan_1931, Aymara_Central, fra_pirotclamer, ozm, mnf, eng_treeoflife, Wolof-NT, deu_meister, Naga_Ao, hun_revised, Faroese, pol_living, krc, cop_bohairicdiacritics, eng_newinternational, Umbundu (011), Sanskrit, Ido, Gujarati, ben_easy, mos_catholic, Gonja, nds, deu_schlachter2000, fra_nouvellesegond, Hindi, Pular (Adlam), Cabecar-NT, gng_2011, Quechua_Cusco, mai, fub, Khün, wmw, fra_paroledevie, Marathi, Mixe_Totontepec, fuv, ben_old, mwm_2010, Crioulo_Upper Guinea (008), Telugu, Sinhala, por_almeidarevista, por_versaofacil, Ndebele, Susu, kab, hbo_leningrad, qub, nin_2011, Tatar, rus_synodal, box, cat, nor, gun, hun_2005, deu_neue, arb, hun_2003, nko, shk, Arabic_Standard, bul, por_almeidaatualizada, Tuareg-PART, spp, Thai (2), khm_standard2005, mar, bzd, ukr_1962, eng_scriptures, Kazakh, Mina, hrv_2000, por_paratodos, gag, isl, mkd, nyy, uzn, kbp, deu_gruenewalder, fra_perret, kmr, swg, Lithuanian, kez, Panjabi_Eastern, sbd, Quichua-NT, bfd, ttq, Mazahua Central, gej, Guaraní_Paraguayan, pol_nowagdansk, cym_colloquial2013, sld, Makonde, mkd_2004, ben_common, bul_modern, tyv, fra_jerusalem2004, Thai, Dinka_Northeastern, bss, por_versaointernacional, Romani_Balkan, Chokwe (Angola), Lamnso', kpj, fra_segond21, ken, urd_revised2010, adj, Nyamwezi, gbo, bud, Zulu-NT, nob_1985, biv, qvw, Bengali, Maithili, Malayalam, swe_b2000, Khmer_Central, mlt, Armenian, lav_1997, Bulgarian, Adyghe, Komi-Permyak, Vietnamese, kin_2001, gsw_alemannisch, nob_2011, kpv, ewe, nan, atg, nnh_2007, kmg, chu, Slovene, Maninkakan_Eastern, Aja, Kpelle_Guinea, Sãotomense, slv, rus_centralasian, lit, nno_2011, Chamorro-PART, Serbian, Ewe-NT, hrv, eng_internationalstandard, ces_kralicka, fra_crampon, kaa, rus_churchslavonic, ttr, Cree_Swampy, tur_southernazeri, Osetin, rus_modern2011, xho_1996, Soninke, Portuguese, Kaonde, ukr_2009, fra_zadockahn, est_portions, gug, tbz, lee, Estonian-PART, mhr, yor, lit_1999, Otomi_Mezquital, guj, Norwegian, Fon, Slovak, Czech, Yukaghir_Northern, Zapotec_Miahuatlán, Mixtec_Metlatónoc, dgi, ces_living, lvs, rus_kulakov, Inuktitut_Eastern Canadian, Tai Dam, Bhojpuri, Evenki, ibo, gkn, hye_eastern, rus_slovozhizny2006, ces_novakarlica, ces_ekumenicky, xal, azb, Jola-Fonyi, Chinantec_Chiltepec, ukr_2007, Chickasaw, ces_preklad, Serer-Sine, Dendi, Magahi, Greek (polytonic), Chinantec_Ojitlán, Shilluk, Urdu, lav_ljd_youversion, Arabic, Kabiyé, ukr_1871, slk_standard, hye_western, Tamazight_Standard Morocan, slk_catholic, Ashéninka Perené, Armenian-PART, Shor, Cashinahua, Waama, mah, Cherokee (uppercase), Kulango_Bouna, Latvian-NT, Waorani, Achuar-Shiwiar (1), srp, kor_latinscript, Ukranian-NT, hbo_aleppo, Bulu, ces_bible21, Cherokee (cased), grc_accented, Siona, Ojibwa_Northwestern, bqc, Karelian, vie_bd2011_youversion, Quichua_Chimborazo Highland, vie_1926compounds, vie_banphothong, Otuho, Cherokee-NT, vie_2002, vie_1926nocompounds, vie_revised2010, Awa-Cuaiquer, ell_hellenic1, ojs, Even, Vietnamese-tok, Ojibwa-NT, iku_2012, Vai, grc_ecumenical, crm, ell_koine1894, Tigrigna, gmv_2011, Amharic, amh, (Yeonbyeon), Nuosu, Korean, kor, Chinese_Hakka, Chinese_Wu, Chinese_Yue, Chinese_Gan, Chinese_Mandarin (Traditional), Chinese_Jinyu, Chinese_Min Nan, Japanese, Chinese_Mandarin (Simplified), kor_1985, Vietnamese (Han nom), Japanese-tok, kor_revised, Chinese, Chinese-tok, cmn_sf_ncv_zefania