| ISO | Name | Train char | Dev char | Eval char | Typ / Tok |
|---|---|---|---|---|---|
| *acu* | Achuar | 1149777 | 136773 | 119849 | $5.902^{-05}$ |
| *afr* | Afrikaans | 3229549 | 413064 | 437532 | $1.985^{-05}$ |
| *agr* | Aguaruna | 991098 | 118726 | 103237 | $6.348^{-05}$ |
| *ake* | Akawaio | 960849 | 113905 | 111793 | $5.141^{-05}$ |
| *alb* | Albanian | 3152312 | 402399 | 427612 | $1.808^{-05}$ |
| *amu* | Amuzgo | 1156128 | 142241 | 132194 | $5.662^{-05}$ |
| *bsn* | Barasana | 1397953 | 171482 | 162042 | $4.736^{-05}$ |
| *cak* | Cakchiquel | 1404839 | 169031 | 161609 | $4.033^{-05}$ |
| *ceb* | Cebuano | 3985326 | 509809 | 536615 | $1.471^{-05}$ |
| *ces* | Czech | 2756308 | 349505 | 371027 | $2.675^{-05}$ |
| *cha* | Chamorro | 641087 | 66469 | 67935 | $1.032^{-04}$ |
| *chq* | Chinantec | 1548993 | 174921 | 164087 | $4.502^{-05}$ |
| *cjp* | Cabecar | 856441 | 100035 | 97246 | $8.256^{-05}$ |
| *cni* | Campa | 1149737 | 133104 | 120600 | $3.990^{-05}$ |
| *dan* | Danish | 2774922 | 364278 | 385352 | $2.298^{-05}$ |
| *deu* | German | 3195266 | 391700 | 417235 | $2.023^{-05}$ |
| *dik* | Dinka | 716411 | 84429 | 81572 | $6.800^{-05}$ |
| *dje* | Zarma | 3126629 | 372921 | 405747 | $1.767^{-05}$ |
| *djk* | Aukan | 1083303 | 129770 | 124682 | $5.083^{-05}$ |
| *dop* | Lukpa | 864347 | 96068 | 95094 | $6.442^{-05}$ |
| *eng* | English | 3238389 | 418697 | 450324 | $1.509^{-05}$ |
| *epo* | Esperanto | 3029361 | 391874 | 409980 | $1.514^{-05}$ |
| *est* | Estonian | 778681 | 177680 | 5329 | $8.007^{-05}$ |
| *eus* | Basque | 801072 | 94904 | 93712 | $7.073^{-05}$ |
| *ewe* | Ewe | 870990 | 97081 | 94738 | $9.315^{-05}$ |
| *fin* | Finnish | 3195802 | 402386 | 426808 | $1.789^{-05}$ |
| *fra* | French | 3246315 | 404464 | 435820 | $2.129^{-05}$ |
| *gbi* | Galela | 1347199 | 144215 | 129606 | $4.442^{-05}$ |
| *gla* | Gaelic | 68110 | 6802 | 7167 | $5.848^{-04}$ |
| *glv* | Manx | 392897 | 58690 | 48239 | $1.360^{-04}$ |
| *hat* | Creole | 3332162 | 400606 | 440487 | $1.989^{-05}$ |
| *hrv* | Croatian | 2594494 | 340781 | 359694 | $2.549^{-05}$ |
| *hun* | Hungarian | 3020721 | 376738 | 408697 | $2.391^{-05}$ |
| *ind* | Indonesian | 3528757 | 405822 | 454277 | $1.823^{-05}$ |
| *isl* | Icelandic | 2968652 | 376562 | 389091 | $2.517^{-05}$ |
| *ita* | Italian | 2979890 | 388587 | 409629 | $2.091^{-05}$ |
| *jak* | Jakalteko | 1116611 | 131793 | 122853 | $5.615^{-05}$ |
| *jiv* | Shuar | 888886 | 98309 | 97483 | $5.624^{-05}$ |
| *kab* | Kabyle | 798503 | 91677 | 87964 | $7.770^{-05}$ |
| *kbh* | Camsa | 1373946 | 144068 | 140039 | $5.307^{-05}$ |
| *kek* | Q'eqchi' | 4375494 | 525831 | 517455 | $1.735^{-05}$ |
| *lat* | Latin | 2700731 | 325553 | 342365 | $1.514^{-05}$ |
| *lav* | Latvian | 644923 | 77572 | 78263 | $1.161^{-04}$ |
| *lit* | Lithuanian | 2531703 | 313391 | 309237 | $2.536^{-05}$ |
| *mam* | Mam | 1053107 | 119781 | 112869 | $6.222^{-05}$ |
| *mri* | Maori | 3504361 | 437499 | 456364 | $1.501^{-05}$ |
| *nhg* | Nahuatl | 1126416 | 135355 | 126213 | $5.548^{-05}$ |
| *nld* | Dutch | 3224058 | 392079 | 416432 | $2.033^{-05}$ |
| *nor* | Norwegian | 2941245 | 374161 | 392574 | $2.508^{-05}$ |
| *pck* | Paite | 3174091 | 404462 | 401042 | $1.784^{-05}$ |
| *plt* | Malagasy | 3744462 | 468678 | 477671 | $1.705^{-05}$ |
| *pol* | Polish | 2963005 | 374471 | 398263 | $2.088^{-05}$ |
| *por* | Portuguese | 3010541 | 380551 | 404559 | $2.450^{-05}$ |
| *pot* | Potawatomi | 212243 | 25336 | 24020 | $1.911^{-04}$ |
| *ppk* | Uma | 1050858 | 115947 | 110127 | $5.090^{-05}$ |
| *quc* | K'iche' | 1153252 | 131623 | 127281 | $5.382^{-05}$ |
| *quw* | Quichua | 792834 | 93791 | 90930 | $8.593^{-05}$ |
| *rom* | Romani | 818094 | 91328 | 89036 | $7.912^{-05}$ |
| *ron* | Romanian | 3107966 | 394280 | 419241 | $1.760^{-05}$ |
| *shi* | Tachelhit | 738833 | 82470 | 79965 | $5.659^{-05}$ |
| *slk* | Slovak | 2821379 | 361684 | 385812 | $2.662^{-05}$ |
| *slv* | Slovene | 2883854 | 369397 | 381124 | $2.366^{-05}$ |
| *sna* | Shona | 3013970 | 384548 | 407708 | $2.154^{-05}$ |
| *som* | Somali | 3750398 | 468849 | 498051 | $1.314^{-05}$ |
| *spa* | Spanish | 3082800 | 388079 | 410641 | $2.190^{-05}$ |
| *srp* | Serbian | 2503088 | 319878 | 341496 | $2.402^{-05}$ |
| *ssw* | Swahili | 763781 | 84855 | 82704 | $6.979^{-05}$ |
| *swe* | Swedish | 3206128 | 403712 | 427604 | $1.932^{-05}$ |
| *tgl* | Tagalog | 3848347 | 477839 | 505730 | $1.283^{-05}$ |
| *tmh* | Tuareg | 270666 | 30146 | 31636 | $2.436^{-04}$ |
| *tur* | Turkish | 2719803 | 318179 | 345666 | $2.601^{-05}$ |
| *usp* | Uspanteco | 1134539 | 131631 | 125891 | $5.747^{-05}$ |
| *vie* | Vietnamese | 3194226 | 379697 | 417496 | $4.635^{-05}$ |
| *wal* | Wolaytta | 837506 | 98141 | 96951 | $6.004^{-05}$ |
| *wol* | Wolof | 683480 | 80261 | 77575 | $8.201^{-05}$ |
| *xho* | Xhosa | 3005476 | 377342 | 399338 | $1.692^{-05}$ |
| *zul* | Zulu | 690644 | 80944 | 77975 | $8.946^{-05}$ |

Table 1: Stats for all languages in our corpus: for each ISO 639-3 code, we list (from left to right) the language name, the character count (in the train, development, and evaluation sets), and the type-to-token ratio.

## Appendix A: list of ISO 639-3 codes and language names

In Tab. 1 the ISO 639-3 codes for each language are associated with the corresponding language name. In addition to this information, Tab. 1 provides the total count of characters for the three data splits, and the type-to-token ratio.

## Appendix B: Typological Features

The 245 binarized typological features from Littell et al. (2017) that characterize the general properties of each language are plotted as a heat map in Fig. 1. Features are related to syntax if their name starts with $S$, to phonology if it starts with $P$, and to phonemic inventories if it starts with $INV$. Note how some values are so rare that they belong exclusively to a single language in the sample, e.g. the vowel /ɔ/ for Thai.

## References

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of EACL*, pages 8–14.

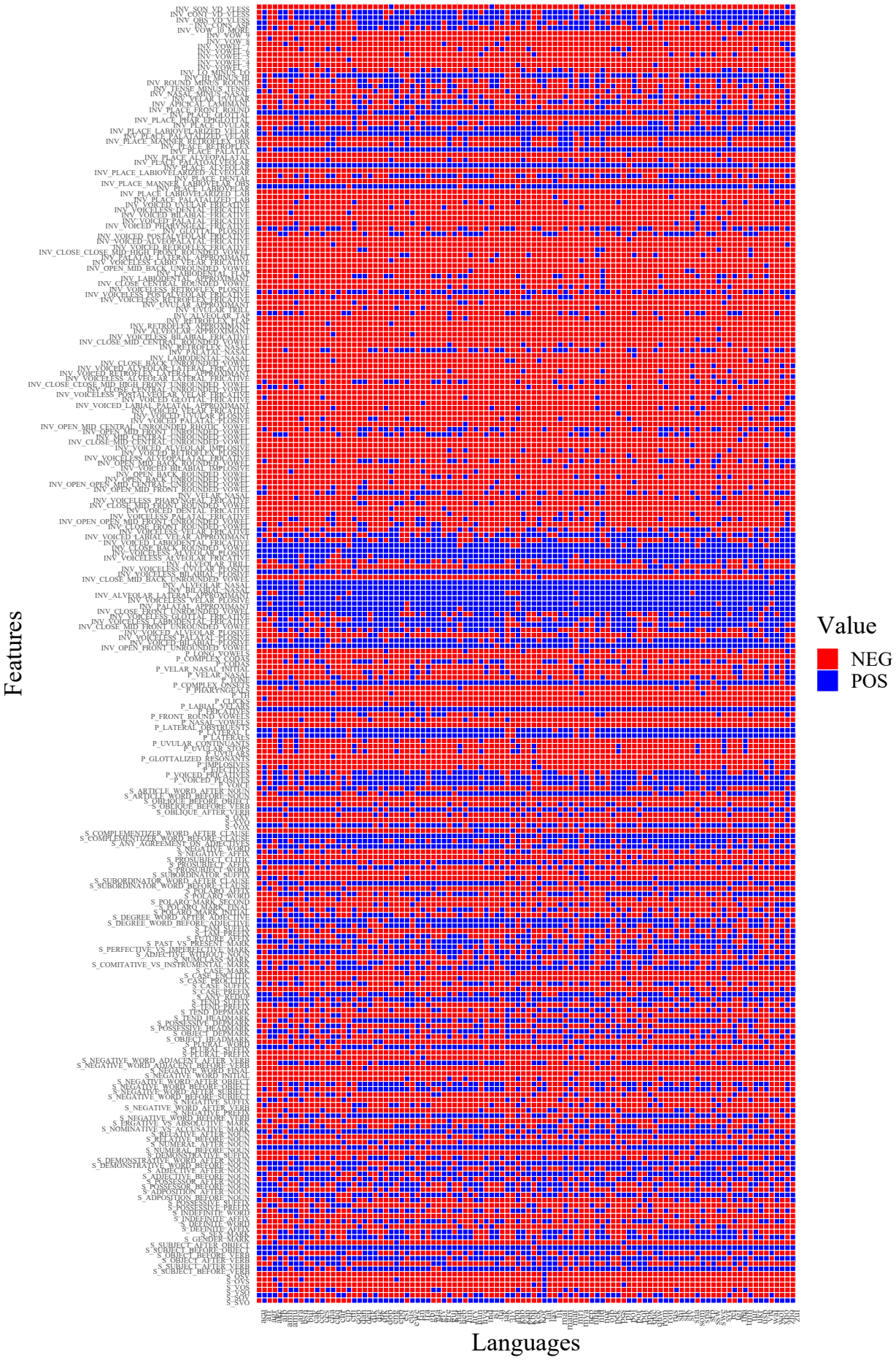


Figure 1: Binary values of the typological features from Littell et al. (2017) (y-axis) for each language (x-axis).