Figure 6.5: Training data statistics for 600 languages, sorted in descending order by English token count, obtained after deduplication and filtering (see Section 6.3.1). The full name for these ISO 639-3 codes can be looked up using MTDATA, e.g. mtdata-iso eng.

Cover 92 (medium-resource) languages, and WMT NewsTests (Bojar et al., 2017a, 2018; Barrault et al., 2019a, 2020) having high quality test sets for 23 (mostly high resource languages).