### 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 *IN*. Note how some values are so rare that they belong exclusively to a single language in the sample, e.g. the vowel /a/ for Thai.

### References

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