Features	Dimension	Mantel r	Mantel p	Mantel z
geography	299	0.647	0.007*	3.443
family	3718	0.329	0.023*	2.711
syntax	103	0.649	0.007*	3.787
inventory	158	0.155	0.459	0.782
phonology	28	0.397	0.046	1.943
Malaviya, Neubig, and Littell (2017)	512	-0.431	0.264	-1.235

Table 9: Mantel test on the correlation between similarity judgments from Multi-SimLex and linguistic features from typological databases.

Oceania (Dryer 2013). In turn, geographical proximity leads to similar judgment patterns, as mentioned above. On the other hand, we find no correlation with phonology and inventory, as expected, nor with the bottom-up typological features from Malaviya, Neubig, and Littell (2017).

6. Cross-Lingual Multi-SimLex Datasets

A crucial advantage of having semantically aligned monolingual datasets across different languages is the potential to create *cross-lingual semantic similarity datasets*. Such datasets allow for probing the quality of cross-lingual representation learning algorithms (Camacho-Collados et al. 2017; Conneau et al. 2018a; Chen and Cardie 2018; Doval et al. 2018; Ruder, Vulić, and Søgaard 2019; Conneau and Lample 2019; Ruder, Søgaard, and Vulić 2019) as an intrinsic evaluation task. However, the cross-lingual datasets previous work relied upon (Camacho-Collados et al. 2017) were limited to a homogeneous set of high-resource languages (e.g., English, German, Italian, Spanish) and a small number of concept pairs (all less than 1K pairs). We address both problems by 1) using a typologically more diverse language sample, and 2) relying on a substantially larger English dataset as a source for the cross-lingual datasets: 1,888 pairs in this work versus 500 pairs in the work of Camacho-Collados et al. (2017). As a result, each of our cross-lingual datasets contains a substantially larger number of concept pairs, as shown in Table 11. The cross-lingual Multi-Simlex datasets are constructed automatically, leveraging word pair translations and annotations collected in all 12 languages. This yields a total of 66 cross-lingual datasets, one for each possible combination of languages. Table 11 provides the final number of concept pairs, which lie between 2,031 and 3,480 pairs for each cross-lingual dataset, whereas Table 10 shows some sample pairs with their corresponding similarity scores.

The automatic creation and verification of cross-lingual datasets closely follows the procedure first outlined by Camacho-Collados, Pilehvar, and Navigli (2015) and later adopted by Camacho-Collados et al. (2017) (for semantic similarity) and Vulić, Ponzetto, and Glavaš (2019) (for graded lexical entailment). First, given two languages, we intersect their aligned concept pairs obtained through translation. For instance, starting from the aligned pairs *attroupement – foule* in French and *rahvasumm – rahvahulk* in Estonian, we construct two cross-lingual pairs *attroupement – rahvaluk* and *rahvasumm – foule*. The scores of cross-lingual pairs are then computed as averages of the two corresponding monolingual scores. Finally, in order to filter out concept pairs whose semantic meaning was not preserved during this operation, we retain only cross-lingual pairs for which the