Language	Word Pair	POS	Rating all participants agree with
ENG	trial – test	N	4-5
SWA	archbishop – bishop	N	4-5
SPA, CYM	start – begin	V	5-6
ENG	smart – intelligent	ADJ	5-6
ENG, SPA	quick – rapid	ADJ	5-6
SPA	circumstance – situation	N	5-6
CYM	football – soccer	N	5-6
SWA	football – soccer	N	6
SWA	pause – wait	V	6
SWA	money – cash	N	6
CYM	friend – buddy	N	6

Table 8: Examples of concept pairs with their similarity scores from four languages where all participants show strong agreement in their rating.

where the '*tyngu*' may not have been a commonly used or even a known word choice for annotators, pointing out potential regional or generational differences in language use.

Table 8 presents examples of concept pairs from English, Spanish, Kiswahili, and Welsh on which the participants agreed the most. For example, in English all participants rated the similarity of *trial – test* to be 4 or 5. In Spanish and Welsh, all participants rated *start – begin* to correspond to a score of 5 or 6. In Kiswahili, *money – cash* received a similarity rating of 6 from every participant. While there are numerous examples of concept pairs in these languages where the participants agreed on a similarity score of 4 or higher, it is worth noting that none of these languages had a single pair where all participants agreed on either 1-2, 2-3, or 3-4 similarity rating. Interestingly, in English all pairs where all the participants agreed on a 5-6 similarity score were adjectives.

5.4 Effect of Language Affinity on Similarity Scores

Based on the analysis in Figure 1 and inspecting the anecdotal examples in the previous section, it is evident that the correlation between similarity scores across languages is not random. To corroborate this intuition, we visualize the vectors of similarity scores for each single language by reducing their dimensionality to 2 via Principal Component Analysis (Pearson 1901). The resulting scatter plot in Figure 3 reveals that languages from the same family or branch have similar patterns in the scores. In particular, Russian and Polish (both Slavic), Finnish and Estonian (both Uralic), Cantonese and Mandarin Chinese (both Sinitic), and Spanish and French (both Romance) are all neighbors.

In order to quantify exactly the effect of language affinity on the similarity scores, we run correlation analyses between these and language features. In particular, we extract feature vectors from URIEL (Littell et al. 2017), a massively multilingual typological database that collects and normalizes information compiled by grammarians and field linguists about the world's languages. In particular, we focus on information about *geography* (the areas where the language speakers are concentrated), *family* (the phylogenetic tree each language belongs to), and typology (including *syntax*, phonological *inventory*, and *phonology*).[5] Moreover, we consider typological representations of languages that are not manually crafted by experts, but rather learned from texts. Malaviya, Neubig, and

5 For the extraction of these features, we employed `lang2vec`: `github.com/antonisa/lang2vec`