Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
Nouns	0.661	0.622	0.659	0.558	0.647	0.698	0.538	0.606	0.524	0.582	0.626	0.727
Adjectives	0.757	0.698	0.823	0.695	0.721	0.741	0.683	0.699	0.625	0.64	0.658	0.785
Verbs	0.694	0.604	0.707	0.58	0.644	0.691	0.615	0.593	0.555	0.588	0.631	0.76
Adverbs	0.699	0.593	0.695	0.579	0.646	0.595	0.561	0.543	0.535	0.563	0.562	0.716
Overall	0.68	0.619	0.698	0.583	0.646	0.697	0.572	0.609	0.53	0.576	0.623	0.733

Table 4: Average pairwise inter-annotator agreement (APIAA). A score of 0.6 and above indicates strong agreement.

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
Nouns	0.757	0.747	0.766	0.696	0.766	0.809	0.68	0.717	0.657	0.71	0.725	0.804
Adjectives	0.800	0.789	0.865	0.79	0.792	0.831	0.754	0.792	0.737	0.743	0.686	0.811
Verbs	0.774	0.733	0.811	0.715	0.757	0.808	0.72	0.722	0.69	0.71	0.702	0.784
Adverbs	0.749	0.693	0.777	0.697	0.748	0.729	0.645	0.655	0.608	0.671	0.623	0.716
Overall	0.764	0.742	0.794	0.715	0.76	0.812	0.699	0.723	0.667	0.703	0.71	0.792

Table 5: Average mean inter-annotator agreement (AMIAA). A score of 0.6 and above indicates strong agreement.

measurements. The languages with the highest annotator agreement were French (FRA) and Yue Chinese (YUE), while Russian (RUS) had the lowest overall IAA scores. These scores, however, are still considered to be 'moderately strong agreement'.

5.3 Data Analysis

Similarity Score Distributions. Across all languages, the average score (mean $= 1.61$, median$= 1.1$) is on the lower side of the similarity scale. However, looking closer at the scores of each language in Table 6, we indicate notable differences in both the averages and the spread of scores. Notably, French has the highest average of similarity scores (mean$= 2.61$, median$= 2.5$), while Kiswahili has the lowest average (mean$= 1.28$, median$= 0.5$). Russian has the lowest spread ($\sigma = 1.37$), while Polish has the largest ($\sigma = 1.62$). All of the languages are strongly correlated with each other, as shown in Figure 1, where all of the Spearman's correlation coefficients are greater than 0.6 for all language pairs. Languages that share the same language family are highly correlated (e.g, CMN-YUE, RUS-POL, EST-FIN). In addition, we observe high correlations between English and most other languages, as expected. This is due to the effect of using English as the base/anchor language to create the dataset. In simple words, if one translates to two languages L_1 and L_2 starting from the same set of pairs in English, it is higly likely that L_1 and L_2 will diverge from English in different ways. Therefore, the similarity between L_1-ENG and L_2-ENG is expected to be higher than between L_1-L_2, especially if L_1 and L_2 are typologically dissimilar languages (e.g., HEB-CMN, see Figure 1). This phenomenon is well documented in related prior work (Leviant and Reichart 2015; Camacho-Collados et al. 2017; Mrkšić et al. 2017; Vulić, Ponzetto, and Glavaš 2019). While we acknowledge this as a slight artifact of the dataset design, it would otherwise be impossible to construct a semantically aligned and comprehensive dataset across a large number of languages.

We also report differences in the distribution of the frequency of words among the languages in Multi-SimLex. Figure 2 shows six example languages, where each bar segment shows the proportion of words in each language that occur in the given frequency range. For example, the 10K-20K segment of the bars represents the proportion