Pair	Concept-1	Concept-2	Score	Pair	Concept-1	Concept-2	Score
CYM-ENG	rhyddid	liberty	5.37	CMN-EST	可能	optimistlikult	0.83
CYM-POL	plentynaidd	niemądry	2.15	FIN-SWA	psykologia	sayansi	2.20
SWA-ENG	kutimiza	accomplish	5.24	ENG-FRA	normally	quotidiennement	2.41
CMN-FRA	有弹性	flexible	4.08	FIN-SPA	auto	bicicleta	0.85
FIN-SPA	tietämättömyys	inteligencia	0.55	CMN-YUE	使灰心	使气馁	4.78
SPA-FRA	ganador	candidat	2.15	CYM-SWA	sefyllfa	mazingira	1.90
EST-YUE	takso	巴士	2.08	EST-SPA	armee	legión	3.25
ENG-FIN	orange	sitrushedelmä	3.43	FIN-EST	halveksuva	põlglik	5.55
SPA-POL	palabra	wskazówka	0.55	CMN-CYM	学生	disgybl	4.45
POL-SWA	prawdopodobnie	uwezekano	4.05	POL-ENG	grawitacja	meteor	0.27

Table 10: Example concept pairs with their scores from a selection of cross-lingual Multi-SimLex datasets.

	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
CMN	1,888	–	–	–	–	–	–	–	–	–	–	–
CYM	3,085	1,888	–	–	–	–	–	–	–	–	–	–
ENG	3,151	3,380	1,888	–	–	–	–	–	–	–	–	–
EST	3,188	3,305	3,364	1,888	–	–	–	–	–	–	–	–
FIN	3,137	3,274	3,352	3,386	1,888	–	–	–	–	–	–	–
FRA	2,243	2,301	2,284	2,787	2,682	1,888	–	–	–	–	–	–
HEB	3,056	3,209	3,274	3,358	3,243	2,903	1,888	–	–	–	–	–
POL	3,009	3,175	3,274	3,310	3,294	2,379	3,201	1,888	–	–	–	–
RUS	3,032	3,196	3,222	3,339	3,257	2,219	3,226	3,209	1,888	–	–	–
SPA	3,116	3,205	3,318	3,312	3,256	2,645	3,256	3,250	3,189	1,888	–	–
SWA	2,807	2,926	2,828	2,845	2,900	2,031	2,775	2,819	2,855	2,811	1,888	–
YUE	3,480	3,062	3,099	3,080	3,063	2,313	3,005	2,950	2,966	3,053	2,821	1,888

Table 11: The sizes of all monolingual (main diagonal) and cross-lingual datasets.

corresponding monolingual scores (s_s, s_t) differ at most by one fifth of the full scale (i.e., $\mid s_s - s_t \mid \leq 1.2$). This heuristic mitigates the noise due to cross-lingual semantic shifts (Camacho-Collados et al. 2017; Vulić, Ponzetto, and Glavaš 2019). We refer the reader to the work of Camacho-Collados, Pilehvar, and Navigli (2015) for a detailed technical description of the procedure.

To assess the quality of the resulting cross-lingual datasets, we have conducted a verification experiment similar to Vulić, Ponzetto, and Glavaš (2019). We randomly sampled 300 concept pairs in the English-Spanish, English-French, and English-Mandarin cross-lingual datasets. Subsequently, we asked bilingual native speakers to provide similarity judgments of each pair. The Spearman's correlation score ρ between automatically induced and manually collected ratings achieves $\rho \geq 0.90$ on all samples, which confirms the viability of the automatic construction procedure.

Score and Class Distributions. The summary of score and class distributions across all 66 cross-lingual datasets are provided in Figure 4a and Figure 4b, respectively. First, it is obvious that the distribution over the four POS classes largely adheres to that of the original monolingual Multi-SimLex datasets, and that the variance is quite low: e.g., the ENG-FRA dataset contains the lowest proportion of nouns (49.21%) and the highest proportion of verbs (27.1%), adjectives (15.28%), and adverbs (8.41%). On the other hand, the distribution over similarity intervals in Figure 4a shows a much greater variance. This is again expected as this pattern resurfaces in monolingual datasets (see Table 6). It is also evident that the data are skewed towards lower-similarity concept pairs. However, due to the joint size of all cross-lingual datasets (see Table 11), even the least represented