Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
R1: Start	13	12	14	12	13	10	11	12	12	12	11	13
R3: End	11	10	13	10	10	10	10	10	10	10	10	11

Table 3: Number of human annotators. R1 = Annotation Round 1, R3 = Round 3.

Round 2: We compare the scores of all annotators and identify the pairs for each annotator that have shown the most disagreement. We ask the annotators to reconsider the assigned scores for those pairs only. The annotators may chose to either change or keep the scores. As in the case with Round 1, the annotators have no access to the scores of the other annotators, and the process is anonymous. This process gives a chance for annotators to correct errors or reconsider their judgments, and has been shown to be very effective in reaching consensus, as reported by Pilehvar et al. (2018). We used a very similar procedure as Pilehvar et al. (2018) to identify the pairs with the most disagreement; for each annotator, we marked the ith pair if the rated score s_i falls within: $s_i \geq \mu_i + 1.5$ or $s_i \leq \mu_i - 1.5$, where μ_i is the mean of the other annotators' scores.

Round 3: We compute the average agreement for each annotator (with the other annotators), by measuring the average Spearman's correlation against all other annotators. We discard the scores of annotators that have shown the least average agreement with all other annotators, while we maintain at least ten annotators per language by the end of this round. The actual process is done in multiple iterations: (S1) we measure the average agreement for each annotator with every other annotator (this corresponds to the APIAA measure, see later); (S2) if we still have more than 10 valid annotators and the lowest average score is higher than in the previous iteration, we remove the lowest one, and rerun S1. Table 3 shows the number of annotators at both the start (Round 1) and end (Round 3) of our process for each language.

We measure the agreement between annotators using two metrics, average pairwise inter-annotator agreement (APIAA), and average mean inter-annotator agreement (AMIAA). Both of these use Spearman's correlation (ρ) between annotators scores, the only difference is how they are averaged. They are computed as follows:

$$1)\text{APIAA} = \frac{2\sum_{i,j}\rho(s_i,s_j)}{N(N-1)} \qquad 2)\text{AMIAA} = \frac{\sum_i \rho(s_i,\mu_i)}{N}\text{, where: } \mu_i = \frac{\sum_{j,j\neq i} s_j}{N-1} \tag{1}$$

where $\rho(s_i, s_j)$ is the Spearman's correlation between annotators i and j's scores (s_i,s_j) for all pairs in the dataset, and N is the number of annotators. APIAA has been used widely as the standard measure for inter-annotator agreement, including in the original SimLex paper (Hill, Reichart, and Korhonen 2015). It simply averages the pairwise Spearman's correlation between all annotators. On the other hand, AMIAA compares the average Spearman's correlation of one held-out annotator with the average of all the other $N-1$ annotators, and then averages across all N 'held-out' annotators. It smooths individual annotator effects and arguably serves as a better upper bound than APIAA (Gerz et al. 2016; Vulić et al. 2017a; Pilehvar et al. 2018, *inter alia*).

We present the respective APIAA and AMIAA scores in Table 4 and Table 5 for all part-of-speech subsets, as well as the agreement for the full datasets. As reported in prior work (Gerz et al. 2016; Vulić et al. 2017a), AMIAA scores are typically higher than APIAA scores. Crucially, the results indicate 'strong agreement' (across all languages) using both