Languages:	CMN	CYM	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE	*Avg*
Nouns	84.5	80.0	90.0	87.3	78.2	98.2	90.0	95.5	85.5	80.0	77.3	86.0
Adjectives	88.5	88.5	61.5	73.1	69.2	100.0	84.6	100.0	69.2	88.5	84.6	82.5
Verbs	88.0	74.0	82.0	76.0	78.0	100.0	74.0	100.0	74.0	76.0	86.0	82.5
Adverbs	92.9	100.0	57.1	78.6	92.9	100.0	85.7	100.0	85.7	85.7	78.6	87.0
Overall	86.5	81.0	82.0	82.0	78.0	99.0	85.0	97.5	80.5	81.0	80.5	84.8

Table 2: Inter-translator agreement (% of matched translated words) by independent translators using a randomly selected 100-pair English sample from the Multi-SimLex dataset, and the corresponding 100-pair samples from the other datasets.

doctor both translate to the same word in Estonian (*arst*); the less formal word *doktor* is used as a translation of *doctor* to generate a valid pair.

We measure the quality of the translated pairs by using a random sample set of 100 pairs (from the 1,888 pairs) to be translated by an independent translator for each target language. The sample is proportionally stratified according to the part-of-speech categories. The independent translator is given identical instructions to the main translator; we then measure the percentage of matched translated words between the two translations of the sample set. Table 2 summarizes the inter-translator agreement results for all languages and by part-of-speech subsets. Overall across all languages, the agreement is 84.8%, which is similar to prior work (Camacho-Collados et al. 2017; Vulić, Ponzetto, and Glavaš 2019).

5.2 Guidelines and Word Pair Scoring

Across all languages, 145 human annotators were asked to score all 1,888 pairs (in their given language). We finally collect at least ten valid annotations for each word pair in each language. All annotators were required to abide by the following instructions:

1. Each annotator must assign an integer score between 0 and 6 (inclusive) indicating how semantically similar the two words in a given pair are. A score of 6 indicates very high similarity (i.e., perfect synonymy), while zero indicates no similarity.

2. Each annotator must score the entire set of 1,888 pairs in the dataset. The pairs must not be shared between different annotators.

3. Annotators are able to break the workload over a period of approximately 2-3 weeks, and are able to use external sources (e.g. dictionaries, thesauri, WordNet) if required.

4. Annotators are kept anonymous, and are not able to communicate with each other during the annotation process.

The selection criteria for the annotators required that all annotators must be native speakers of the target language. Preference to annotators with university education was given, but not required. Annotators were asked to complete a spreadsheet containing the translated pairs of words, as well as the part-of-speech, and a column to enter the score. The annotators did not have access to the original pairs in English.

To ensure the quality of the collected ratings, we have employed an *adjudication protocol* similar to the one proposed and validated by Pilehvar et al. (2018). It consists of the following three rounds:

Round 1: All annotators are asked to follow the instructions outlined above, and to rate all 1,888 pairs with integer scores between 0 and 6.