	CMN- ENG	ENG- FRA	ENG- SPA	ENG- RUS	EST- FIN	EST- HEB	FIN- HEB	FRA- SPA	POL- RUS	POL- SPA
UNSUPER	.565	.662	.498	.511	.510	.465	.445	.600	.390	.398
SUPER (1k)	.575	.602	.453	.376	.378	.363	.442	.588	.399	.406
+SL (1k)	.577	.703	.547	.548	**.591**	.513	.488	.639	**.439**	.456
SUPER (5k)	**.587**	.704	.542	.535	.518	.473	.585	.631	.455	.463
+SL (5k)	.581	**.707**	**.548**	**.551**	.556	**.525**	**.589**	**.645**	.432	**.476**

Table 16: Results on a selection of cross-lingual Multi-SimLex datasets where the fully unsupervised (UNSUPER) CLWE variant yields reasonable performance. We also show the results with supervised VECMAP without self-learning (SUPER) and with self-learning (+SL), with two seed dictionary sizes: 1k and 5k pairs; see §8.1 for more detail. Highest scores for each language pair are in **bold**.

	CMN-FIN	CMN-RUS	CMN-YUE	CYM-FIN	CYM-FRA	CYM-POL	FIN-SWA
UNSUPER	.049	.032	.004	.020	.015	.028	.013
SUPER (1k)	.410	.388	.372	.384	.475	.326	.206
+SL (1k)	**.590**	**.537**	**.458**	**.471**	**.578**	**.380**	**.264**

Table 17: Results on a selection of cross-lingual Multi-SimLex datasets where the fully unsupervised (UNSUPER) CLWE variant fails to learn a coherent shared cross-lingual space. See also the caption of Table 16.

multilingual model is trained on a much smaller set of languages (17 versus 100), and further improvements can be achieved by training a dedicated bilingual model. Finally, leveraging bilingual parallel data seems to offer additional slight gains, but a tiny difference between XLM-2 and XLM-2++ also suggests that this rich bilingual information is not used in the optimal way within the XLM architecture for semantic similarity.

In summary, these results indicate that, in order to improve performance in cross-lingual transfer tasks, more work should be invested into 1) pretraining dedicated language pair-specific models, and 2) creative ways of leveraging available cross-lingual supervision (e.g., word translation pairs, parallel or comparable corpora) (Liu et al. 2019a; Wu et al. 2019; Cao, Kitaev, and Klein 2020) with pretraining paradigms such as BERT and XLM. Using such cross-lingual supervision could lead to similar benefits as indicated by the results obtained with static cross-lingual word embeddings (see Table 16 and Table 17). We believe that Multi-SimLex can serve as a valuable means to track and guide future progress in this research area.

9. Conclusion and Future Work

We have presented Multi-SimLex, a resource containing human judgments on the semantic similarity of word pairs for 12 monolingual and 66 cross-lingual datasets. The languages covered are typologically diverse and include also under-resourced ones, such as Welsh and Kiswahili. The resource covers an unprecedented amount of 1,888 word pairs, carefully balanced according to their similarity score, frequency, concreteness, part-of-speech class, and lexical field. In addition to Multi-Simlex, we release the detailed protocol we followed to create this resource. We hope that our consistent guidelines will encourage researchers to translate and annotate Multi-Simlex -style datasets for