	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
CMN		.076	.348	.139	.154	.392	.190	.207	.227	.300	.049	.484
CYM	.041		.087	.017	.049	.095	.033	.072	.085	.089	.002	.083
ENG	.565	.004		.168	.159	.401	.171	.182	.236	.309	.014	.357
EST	.014	.097	.335		.143	.161	.100	.113	.083	.134	.025	.124
FIN	.049	.020	.542	.530		.195	.077	.110	.111	.157	.029	.167
FRA	.224	.015	.662	.559	.533		.191	.229	.297	.382	.038	.382
HEB	.202	.110	.516	.465	.445	.469		.095	.154	.181	.038	.185
POL	.121	.028	.464	.415	.465	.534	.412		.139	.183	.013	.205
RUS	.032	.037	.511	.408	.476	.529	.430	.390		.248	.037	.226
SPA	.546	.048	.498	.450	.490	.600	.462	.398	.419		.055	.313
SWA	-.01	.116	.029	.006	.013	-.05	.033	.052	.035	.045		.043
YUE	.004	.047	.059	.004	.002	.059	.001	.074	.032	.089	-.02	

Table 15: Spearman's ρ correlation scores on all 66 cross-lingual datasets. 1) The scores **below the main diagonal** are computed based on cross-lingual word embeddings (CLWEs) induced by aligning CC+Wiki FT in all languages (except for YUE where we use Wiki FT) in a fully unsupervised way (i.e., without any bilingual supervision). We rely on a standard CLWE mapping-based (i.e., alignment) approach: VECMAP (Artetxe, Labaka, and Agirre 2018b). 2) The scores **above the main diagonal** are computed by obtaining 768-dimensional word-level vectors from pretrained multilingual BERT (M-BERT) following the procedure described in §7.1. For both fully unsupervised VECMAP and M-BERT, we report the results with unsupervised postprocessing enabled: all 2×66 reported scores are obtained using the +ABBT (-10) variant.

taken directly from prior work (Vulić et al. 2019),[21] or extracted from PanLex following the same procedure as in the prior work.

Contextualized Cross-Lingual Word Embeddings. We again evaluate the capacity of (massively) multilingual pretrained language models, M-BERT and XLM-100, to reason over cross-lingual lexical similarity. Implicitly, such an evaluation also evaluates "the intrinsic quality" of shared cross-lingual word-level vector spaces induced by these methods, and their ability to boost cross-lingual transfer between different language pairs. We rely on the same procedure of aggregating the models' subword-level parameters into word-level representations, already described in §7.1.

As in monolingual settings, we can apply unsupervised post-processing steps such as ABTT to both static and contextualized cross-lingual word embeddings.

8.2 Results and Discussion

Main Results and Differences across Language Pairs. A summary of the results on the 66 cross-lingual Multi-SimLex datasets are provided in Table 15 and Figure 6a. The findings confirm several interesting findings from our previous monolingual experiments (§7.2), and also corroborate several hypotheses and findings from prior work, now on a large sample of language pairs and for the task of cross-lingual semantic similarity.

First, we observe that the fully unsupervised VECMAP model, despite being the most robust fully unsupervised method at present, fails to produce a meaningful cross-lingual word vector space for a large number of language pairs (see the bottom triangle

[21] https://github.com/cambridgeltl/panlex-bli