

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
FASTTEXT (CC+Wiki)	FT:INIT											
NOUNS (1,051)	.561	.497	.592	.627	.709	.641	.560	.538	.526	.583	.544	.426
VERBS (469)	.511	.265	.408	.379	.527	.551	.458	.384	.464	.499	.391	.252
ADJ (245)	.448	.338	.564	.401	.546	.616	.467	.284	.349	.401	.344	.288
ADV (123)	.622	.187	.482	.378	.547	.648	.491	.266	.514	.423	.172	.103
FASTTEXT (CC+Wiki)	FT:+ABTT (-3)											
NOUNS	.601	.512	.599	.621	.730	.653	.592	.585	.578	.605	.553	.431
VERBS	.583	.305	.454	.379	.575	.602	.520	.390	.475	.526	.381	.314
ADJ	.526	.372	.601	.427	.592	.646	.483	.316	.409	.411	.402	.312
ADV	.675	.150	.504	.397	.546	.695	.491	.230	.495	.416	.223	.081
M-BERT	M-BERT:+ABTT (-3)											
NOUNS	.517	.091	.446	.191	.210	.364	.191	.188	.266	.418	.142	.539
VERBS	.511	.005	.200	.039	.077	.248	.038	.107	.181	.266	.091	.503
ADJ	.227	.050	.226	.028	.128	.193	.044	.046	.002	.099	.192	.267
ADV	.282	.012	.343	.112	.173	.390	.326	.036	.046	.207	.161	.049
XLNet-100	XLNet:+ABTT (-3)											
ALL	.498	.096	.270	.118	.203	.234	.195	.106	.170	.289	.130	.506
NOUNS	.551	.132	.381	.193	.238	.234	.242	.184	.292	.378	.165	.559
VERBS	.544	.038	.169	.006	.190	.132	.136	.073	.095	.243	.047	.570
ADJ	.356	.140	.256	.081	.179	.185	.150	.046	.022	.100	.220	.291
ADV	.284	.017	.040	.086	.043	.027	.221	.014	.022	.315	.095	.156

Table 13: Spearman’s ρ correlation scores over the four POS classes represented in Multi-SimLex datasets. In addition to the word vectors considered earlier in Table 12, we also report scores for another contextualized model, XLNet-100. The numbers in parentheses refer to the total number of POS-class pairs in the original ENG dataset and, consequently, in all other monolingual datasets.

processing is universally useful, and can lead to huge improvements in correlation scores for many languages. In what follows, we also delve deeper into these analyses.

Impact of Unsupervised Post-Processing. First, the results in Table 12 suggest that applying dimension-wise mean centering to the initial vector spaces has positive impact on word similarity scores in all test languages and for all models, both static and contextualized (see the +MC rows in Table 12). Mimno and Thompson (2017) show that distributional word vectors have a tendency towards narrow clusters in the vector space (i.e., they occupy a narrow cone in the vector space and are therefore anisotropic (Mu, Bhat, and Viswanath 2018; Ethayarajh 2019)), and are prone to the undesired effect of hubness (Radovanović, Nanopoulos, and Ivanović 2010; Lazaridou, Dinu, and Baroni 2015).¹⁵ Applying dimension-wise mean centering has the effect of spreading the vectors across the hyper-plane and mitigating the hubness issue, which consequently improves word-level similarity, as it emerges from the reported results. Previous work has already validated the importance of mean centering for clustering-based tasks (Suzuki et al. 2013), bilingual lexicon induction with cross-lingual word embeddings (Artetxe, Labaka, and Agirre 2018a; Zhang et al. 2019; Vulić et al. 2019), and for modeling lexical semantic change (Schlechtweg et al. 2019). However, to the best of our knowledge, the results

¹⁵ Hubness can be defined as the tendency of some points/vectors (i.e., “hubs”) to be nearest neighbors of many points in a high-dimensional (vector) space (Radovanović, Nanopoulos, and Ivanović 2010; Lazaridou, Dinu, and Baroni 2015; Conneau et al. 2018a)