Manual Clustering and Spatial Arrangement of Verbs for Multilingual Evaluation and Typology Analysis

Olga Majewska, Ivan Vulić, Diana McCarthy, Anna Korhonen
Language Technology Lab, Theoretical and Applied Linguistics, University of Cambridge
{om304, iv250, alk23}@cam.ac.uk, diana@dianamccarthy.co.uk

Abstract

We present the first evaluation of the applicability of a spatial arrangement method (SpAM) to a typologically diverse language sample, and its potential to produce semantic evaluation resources to support multilingual NLP, with a focus on verb semantics. We demonstrate SpAM's utility in allowing for quick bottom-up creation of large-scale evaluation datasets that balance cross-lingual alignment with language specificity. Starting from a shared sample of 825 English verbs, translated into Chinese, Japanese, Finnish, Polish, and Italian, we apply a two-phase annotation process which produces (i) semantic verb classes and (ii) fine-grained similarity scores for nearly 130 thousand verb pairs. We use the two types of verb data to (a) examine cross-lingual similarities and variation, and (b) evaluate the capacity of static and contextualised representation models to accurately reflect verb semantics, contrasting the performance of large language-specific pretraining models with their multilingual equivalent on semantic clustering and lexical similarity, across different domains of verb meaning. We release the data from both phases as a large-scale multilingual resource, comprising 85 verb classes and nearly 130k pairwise similarity scores, offering a wealth of possibilities for further evaluation and research on multilingual verb semantics.

1 Introduction

Many recent efforts in semantic modeling have focused on unsupervised pretraining to extend the benefits offered by recently proposed text encoders (Devlin et al., 2019) to new languages and domains. In these approaches, general language representations are learned from large volumes of unlabeled text, and subsequently leveraged in downstream systems by means of fine-tuning on a given supervised task. The release of large multilingual pretrained encoders (Devlin et al., 2019; Conneau and Lample, 2019) boosted the state of the art on a range of multilingual tasks (Kondratyuk and Straka, 2019; Wang et al., 2019; Pires et al., 2019; Wu and Dredze, 2019; Hu et al., 2020; Artetxe et al., 2020; Qiu et al., 2020; Mueller et al., 2020). In parallel, the number of language-specific pretrained architectures available has also been steadily growing, with the advantage of being more attuned to the properties of the language in question (Virtanen et al., 2019; Nozza et al., 2020). The ease of incorporating these powerful encoders into downstream task pipelines has made them widely popular. However, there is a disproportionate shortage of resources allowing for probing of the learned representations in most languages. The aim of this work is to address this deficit by releasing a *multilingual resource targeting verb semantics* in a typologically diverse selection of languages where no such datasets have hitherto been available. The motivation behind the specific focus on verbs is twofold: (i) the importance of accurate and nuanced representation of verb meaning in light of their pivotal role in sentence structure and the still subpar verbal reasoning ability of SOTA models (Rogers et al., 2020), and (ii) the scarcity of verb data in evaluation datasets currently available. To this end, we employ a recently proposed two-phase data collection method (Majewska et al., 2020) combining semantic clustering (Phase 1) and finer-grained spatial arrangements of words based on their similarity (Phase 2), and evaluate its cross-lingual applicability. Using cross-lingual mappings,

Proceedings of the 28th International Conference on Computational Linguistics, pages 4810–4824
Barcelona, Spain (Online), December 8-13, 2020

Language	ID	N verbs	N classes	N pairs	THR pairs
Chinese Mandarin	ZH	771	17	23990	1898
Finnish	FI	761	16	28641	10065
Italian	IT	817	17	24747	6436
Japanese	JA	704	17	22915	7916
Polish	PL	850	18	28895	6735

Table 1: Data statistics including the number of unique verbs in each sample (translated from English) (**N verbs**), the number of Phase 1 classes (**N classes**), the total number of pairwise scores in the final dataset (**N pairs**) and the thresholded subset of each dataset (**THR pairs**) (See §4.2).

we carry out analyses of cross-language overlap in the semantic classes created in Phase 1, as well as quantitative and qualitative comparisons of the semantic distance matrices from Phase 2. Subsequently, we perform evaluation of static and contextualised representation models on the tasks of lexical similarity and semantic clustering using the data from both phases. This allows us to identify models' strengths and shortcomings, as well as specific challenges posed by the languages' properties and different domains of verb meaning. The collected data, comprising semantic classes and fine-grained pairwise similarity scores for Chinese, Japanese, Finnish, Italian, and Polish, are made freely available with this paper at `https://github.com/om304/Multi-SpA-Verb`.

2 Background and Design Motivation

Word similarity has been widely used as a go-to intrinsic evaluation task, in which rankings of similarity scores computed between word embeddings produced by representation models are compared against ranked human similarity judgments. The dataset design involving sets of word pairs and their associated rating on a discrete scale has been particularly common, due to its reliance on non-expert native speaker judgments, quicker and cheaper to obtain than the large expert-curated lexical-semantic or semantic-syntactic resources such as WordNet (Fellbaum, 1998) or VerbNet (Kipper Schuler, 2005; Kipper et al., 2006). In English, examples include WordSim-353 (Finkelstein et al., 2002; Agirre et al., 2009), MEN (Bruni et al., 2014) and SimLex-999 (Hill et al., 2015). Analogous datasets have been created in other languages, either through translation from an existing English dataset (e.g., from SimLex: German, Italian, and Russian (Leviant and Reichart, 2015), Hebrew and Croatian (Mrkšić et al., 2017) and Polish (Mykowiecka et al., 2018)), or from a new set of concept pairs (e.g., Turkish (Ercan and Yıldız, 2018), Mandarin Chinese (Huang et al., 2019), Japanese (Sakaizawa and Komachi, 2018)). While these datasets are dominated by nouns (e.g., SimLex includes 222 verb pairs), verb-oriented datasets are harder to come by. In English, these include datasets of Yang and Powers (2006) (130 verb pairs), Baker et al. (2014) (143 verb pairs), Gerz et al. (2016) (3,500 verb pairs). A recent multilingual word similarity dataset, Multi-SimLex (Vulić et al., 2020), extends coverage of verb semantic similarity to 469 verb pairs in 12 languages, including Mandarin Chinese, Finnish, and Polish. Another recently introduced large-scale English verb resource of Majewska et al. (2020) (hereafter SpA-Verb) comprises verb classes and unmatched coverage of nearly 30k verb similarity scores. In this work, we demonstrate that their large-scale dataset creation methodology based on spatial arrangement (SpAM) can be extended to other and typologically diverse languages such as Mandarin Chinese, Japanese, Finnish, Polish, and Italian. For each language, we create a dataset comprising 16-18 verb classes with similarity scores between all class members, resulting in over 20k verb pair similarity scores within each language (Table 1).

We start from the English SpA-Verb sample translated into five target languages and apply the two-phase annotation method combining semantic clustering and spatial arrangements based on semantic similarity proposed in Majewska et al. (2020). The method adapts a SpAM approach previously used in cognitive science and psychology in behavioural studies of visual similarity between concrete objects (Kriegeskorte and Mur, 2012; Mur et al., 2013; Cichy et al., 2019) to lexical stimuli. In Phase 1, a large word sample is divided into a number of broad categories of similar and related items. Each of these classes is then used as input in Phase 2, where the related class members are arranged in a 2D space based on their semantic similarity, with similar words placed closer together. Each item placement simultaneously communicates

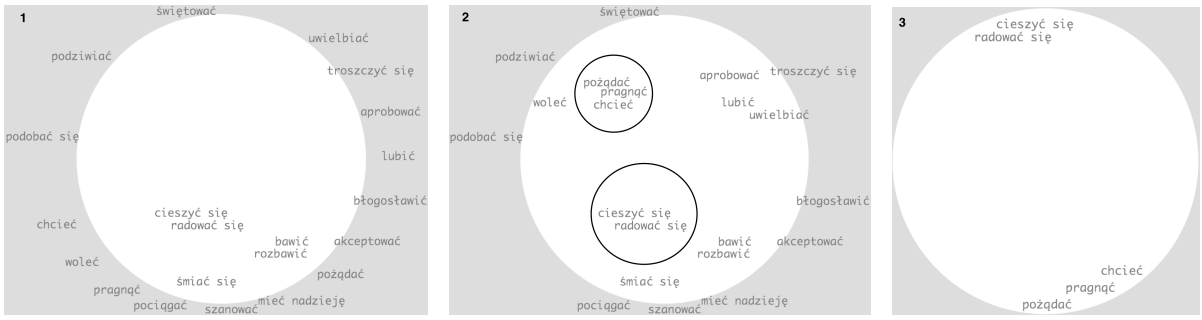

Figure 1: Consecutive Phase 2 trials on a class of Polish emotion verbs. In the first trial (1-2), the whole class is displayed around the arena and word labels are placed one by one based on the similarity of their meaning. Words put closer together in the first trial (2) are subsampled for the subsequent trial (3), and arranged again in a less crowded space (annotators are asked to use the entire space available in each trial and the relative inter-item distances, not the absolute on-screen distances, represent the dissimilarities).

its semantic distance to all other items present and the inter-stimulus Euclidean distances represent pairwise dissimilarities between words in the sample. The arrangements are performed repeatedly over numerous trials first on the entire word set and subsequently on subsets of items, selected by an adaptive algorithm which optimises the evidence collected for the dissimilarity estimates (see Figure 1). The final representational dissimilarity matrix (RDM) estimate is produced by statistically combining the evidence from multiple subsequent 2D arrangements and contains a dissimilarity estimate for each pairing of words in the set (see Kriegeskorte and Mur (2012) for the details). The dissimilarities collected for each Phase 1 class are then normalised to ensure inter-class consistency in the final dataset.

The main advantages of the spatial arrangement method lie in its intuitiveness, rooted in psychology (Lakoff and Johnson, 1999; Gärdenfors, 2004; Casasanto, 2008), and flexibility, due to the reliance on fluid item placements simultaneously expressing multi-way similarity judgments, rather than discrete numerical scores. By repeatedly considering subsets of items, the users reflect on relative differences in meaning between different configurations of words, which decreases bias from placement error, order of presentation and judgment context. The two-phase design offers a practical advantage for porting the method to other languages. The approach starts from a verb sample, rather than a set of word pairs, which allows for easy translation into the target language, avoiding many of the complications encountered in translation of pairs, including cases where both words in the source language pair translate into the same word (e.g., *cup - mug* → Italian *tazza - tazza*), or several pairs in the source language translate into identical target pairs (e.g., *easy - hard*, *easy - difficult* → Polish *łatwy - trudny*).[1]

3 Data Collection and Analysis

We sampled languages from 5 different language families to ensure typological diversity: Sino-Tibetan (Mandarin Chinese ZH), Japonic (Japanese JA), Uralic (Finnish FI), Slavic (Polish PL) and Romance (Italian IT). Following translation from English (EN), the two data collection phases were set up on an online platform (`meadows-research.com`) as two separate studies for each language. Recruitment was carried out on a crowd-sourcing website, `prolific.co`. Participants were native speakers of the target language with at least undergraduate education level and at least a 90% approval rating. Each phase featured a short qualification task testing the participants' understanding of the guidelines.

3.1 Word Sample Translation

Translation was carried out by one native speaker translator per language. In case several equally suitable candidates were identified for one source word, all of them were kept. This was especially true for polysemous English verbs which translated to more than one target verb, each expressing a distinct sense

[1]By translating on a word-by-word basis, each unique source word receives its best target translation, unless no equivalent exists; conversely, if a source word translates into several equally adequate target words, all candidates are included, and thus shortages in one lexical area are compensated in another avoiding major reduction in dataset size.

of the source word (e.g., *bear* → Finnish (1) *kantaa*, 'carry', (2) *sietää*, 'endure'). On the other hand, if two English words had only one adequate translation equivalent, the two-to-one mapping was kept where unavoidable (e.g., *restrict, limit* → Mandarin Chinese 限制). Table 1 shows the number of unique verbs in the final target sample for each language. Additional design choices concerned the following: (i) *multi-word expressions*, which we permitted if they were the natural translation choice, so as to accurately reflect target language lexical semantics; (ii) *intransitive and transitive translation variants* of the same English verb (both variants were kept only if they captured important meaning distinctions beyond valency, e.g., in Polish: *impose*→(1) *narzucać*, 'to force someone to accept something', (2) *narzucać się*, 'to cause inconvenience to someone by demanding their attention'); (iii) *verbal aspect* (translators selected the variant most closely capturing source word meaning, e.g., in Finnish: *jump*→*hypätä* but *bounce*→*hyppiä* (continuative aspect)).

3.2 Phase 1: Semantic Clustering

Five native speakers per language independently performed a rough clustering of the initial verb sample into broad semantic classes. Users dragged words one by one from a queue and placed them in circles representing broad semantic groupings (see Figure 2).

The annotators were instructed to create groupings of similar and related verbs, each containing roughly 30-50 words. This rule of thumb, applied previously (Majewska et al., 2020), ensures similar granularity across languages. To ensure annotation quality, the produced classifications were manually reviewed to identify rogue annotators and low-effort responses (e.g., multiple consecutive words in the queue were placed in the same class indiscriminately or large numbers of words were placed in the trash circle and missing from the final classification), which were subsequently discarded. The final sets of classes for Phase 2 were produced in each language by first identifying the overlap in Phase 1 classifications (i.e., all the verb pairs put in the same class by all annotators in a given language), which determined the class structure and broad semantics of each class (e.g., movement, emotion, communication), and then populating the classes based on majority decisions. Finally, for each language, the cross-subject classes were reviewed manually by a native speaker adjudicator; in the process, the verbs missing from the intersection of individual clusterings were added to valid classes of related verbs (based on the criterion of semantic similarity and relatedness, ensuring semantic coherence of the resultant classes). Phase 1 produced 16-18 classes in each language (Table 1) and took between 2.5 (Finnish) and 3.5 hours (Mandarin Chinese) to complete.

Figure 2: Finnish Phase 1 task interface (zoomed in; the label font is enlarged).

Cross-lingual Overlap. Table 3 summarises Phase 1 output. Given the similar granularity of classifications, we aligned classes with most overlap (via English mappings) and shared broad semantics for an easier comparison. We see a lot of high-level category overlap (e.g., 'possession', 'motion', 'cognition').

To measure the degree of alignment, we calculate pairwise item-level overlap using the B-Cubed metric (Jurgens and Klapaftis, 2013; Amigó et al., 2009) between all language pairs. We examine whether stronger alignment corresponds to a greater typological affinity by confronting the results with the degree of overlap in syntactic, morphological and lexical typological features from the WALS database (Dryer and Haspelmath, 2013) (Table 2).[2] The two languages with the strongest class alignment, Italian and Polish (0.565), also share the most structural properties. Japanese, the only SOV language in the selection, has

	ZH	JA	PL	FI	IT
ZH		0.561	0.526	0.544	0.491
JA	0.304		0.368	0.368	0.246
PL	0.400	0.334		0.684	0.737
FI	0.343	0.332	0.398		0.614
IT	0.397	0.339	0.565	0.399	

Table 2: Pairwise Phase 1 overlap (B-Cubed F-scores) (**lower**△) vs. proportion of shared WALS typological features (**upper**▽).

[2]Feature overlap is a proportion of shared feature values (see Appendix C for a full list of typological features considered).

ID #	Class Label	Chinese size	ρ	Japanese size	ρ	Polish size	ρ	Finnish size	ρ	Italian size	ρ
1	emotion	26	0.38	54	0.66	21	0.36	37	0.57	41	0.63
2	cooking	30	0.39	22	0.53	54	0.30	48	0.42	34	0.39
3	possession	30	0.39	49	0.13	46	0.35	36	0.14	38	0.28
4	motion (S)	74	0.13	76	0.18	92	0.13	87	0.18	86	0.14
5	motion (A/P)	66	0.09	39	0.46	88	0.16	82	0.13	82	0.13
6	sensory perception	32	0.28	-	-	32	0.36	↓	↓	38	0.40
7	physiology	52	0.24	53	0.25	55	0.20	64	0.23	49	0.39
8	state of being	↑	↑	24	0.43	↑	↑	↓	↓	↑	↑
9	change	38	0.44	45	0.35	29	0.47	47	0.26	23	0.58
10	cognition	44	0.24	58	0.29	79	0.17	61	0.18	62	0.24
11	physical contact	47	0.24	↓	↓	55	0.34	37	0.31	44	0.45
12	violence	↑	↑	84	0.31	↑	↑	36	0.40	↑	↑
13	law/crime	75	0.04	↑	↑	68	0.20	69	0.23	73	0.23
14	negative interaction	73	0.07	-	-	50	0.20	50	0.36	↑	↑
15	interaction	69	0.18	60	0.21	49	0.30	79	0.35	69	0.28
16	work/organisation	74	0.07	64	0.24	67	0.19	71	0.21	58	0.38
17	handicraft	51	0.11	63	0.23	60	0.14	↑	↑	71	0.14
18	destruction	39	0.24	46	0.22	39	0.20	52	0.21	26	0.39
19	sound	48	0.13	28	0.32	32	0.34	74	0.33	27	0.25
20	communication	↑	↑	52	0.26	55	0.23	↑	↑	44	0.29
21	combining	-	-	29	0.50	-	-	-	-	-	-

Table 3: Semantic classes produced in Phase 1, aligned cross-lingually based on member overlap (*size* = number of verbs in class, ρ = Spearman's IAA); English labels serve to identify broad semantic categories. ↑/↓ indicate a category is subsumed by the one above or below. S/A/P labels signal arguments typically selected by class members (agent-like (A), patient-like (P), or sole argument of an intransitive verb (S)).

the lowest average pairwise overlap with other languages both in terms of features and Phase 1 classes. Manual examination of the classes provides additional insights into the factors (beyond purely semantic) impacting classification decisions across languages. For instance, in both Polish and Italian we observe a class split corresponding to the reflexive vs. non-reflexive distinction: reflexive motion verbs (e.g., PL *kołysać się*, 'to sway', *obracać się*, 'to spin'; IT *abbassarsi*, 'to lower', *ritirarsi*, 'to retreat') end up separated from their non-reflexive counterparts (PL *kołysać*, *obracać*, IT *abbassare*, *ritirare*). Whereas in Chinese, we observe complex causative verbs (formed with the causative 使 *shǐ*, 'to make, cause') clustered together (e.g., 使失望 *shǐ shīwàng*, 'disappoint', 使心烦 *shǐ xīnfán*, 'upset', 使厌恶 *shǐ yànwù*, 'disgust'), forming a grouping of verbs denoting causing negative emotions. In §3.3, we zoom into specific semantic classes to analyse patterns of similarity and variation in-depth.

3.3 Phase 2: Similarity Multi-Arrangement

The classes from Phase 1 were fed into Phase 2, divided into 5-6 batches of 3-4 classes each. Verbs from one class are annotated independently from all others. Annotators are instructed to arrange presented verbs in a circular arena based on similarity of their meaning, putting similar words closer, disregarding similarity of sound, letters or simple association. For each batch, the aim was to obtain at least 5 valid sets of annotations and recruitment continued until this condition was satisfied. We employed the following post-processing quality assurance protocol: first, we filtered annotators who performed the first arrangement too quickly (i.e., averaging less than 1 sec per word placement upon first seeing the sample, following Majewska et al. (2020)); next, for each class, we filtered out annotators for whom the average pairwise Spearman's correlation of arena dissimilarities with those of all other annotators was more than one standard deviation below the mean of all such average correlations (as done by Hill et al. (2015) and Majewska et al. (2020)). To produce the final dataset and ensure consistency between differently sized classes, we calculated the average of the Euclidean distances from all accepted annotators for each verb pair and then normalised them, using the approach from previous work (Kriegeskorte and Mur, 2012; Majewska et al., 2020) where each dissimilarity matrix is scaled to have a root mean square (RMS) of 1.

Cross-lingual Comparisons. We compute inter-annotator agreement (IAA) in Phase 2 using Spearman's rank correlation coefficient (ρ) as the average correlation of an individual annotator with the average of all other annotators for each class in each language (Table 3). We observe that certain classes proved consistently easier to judge across languages ('emotion', 'change', 'cooking'), while some were

consistently more challenging ('motion', 'handicraft', 'law/crime').[3]

The availability of complete dissimilarity matrices enables analyses of cross-lingual similarities in how concepts pertaining to a given domain are organised. To illustrate this, we focus on two semantic areas, verbs of motion (#4) and change (#9), and compute the correlation between the intersection of distance matrices for all language pairs, and additionally English SpA-Verb data, using the non-parametric Mantel test (Mantel, 1967). We find statistically significant correlations between all pairings of languages ($p \leq .001$), but the results show cross-lingual and cross-domain differences (Table 4). Overall, we observe substantially higher correlations on verbs of change than movement verbs, mirroring the intralingual IAA patterns (Table 3): while there is more room for variation in pairwise distances in a more populated 'motion' class, the alignment on verbs of change is also due to the nature of the class, dominated by antonymous verb pairs of opposite polarity (e.g., *increase-decrease, grow-shrink*), which are consistently spread out in the arena. The moderate to strong correlations recorded indicate that the dimensions which underlie the organisation of concepts in this class - especially the polarity dimension - are cross-lingually shared. As observed in Phase 1 (Table 2), Italian and Polish correlate the most, while Japanese is the least aligned with other languages.

	EN	ZH	JA	PL	FI	IT
EN		0.326	0.364	0.246	0.393	0.434
ZH	0.852		0.373	0.284	0.323	0.259
JA	0.620	0.796		0.311	0.403	0.420
PL	0.729	0.819	0.794		0.234	0.252
FI	0.732	0.725	0.649	0.781		0.430
IT	0.872	0.811	0.684	0.878	0.664	

Table 4: Mantel test results (Pearson's r) for cross-lingual pairs of Phase 2 distance matrices (including EN SpA-Verb) for classes 'motion (S)' (**upper**$\triangledown$) and 'change' (**lower**$\triangle$) (all correlations with $p \leq .001$).

Comparison with the 'motion' class illustrates that there is variation in patterns of cross-lingual affinities across different semantic domains. While Italian correlates the most with English, the correlation with Polish motion verbs is weak. Running agglomerative clustering on top of distance matrices revealed that in all three there emerge subclusters corresponding to the different medium of movement ([*dive, swim, flow*], [*run, walk, crawl*]) and a separation between static and dynamic verbs ([*lounge, poise, remain*], [*chase, dance, dash*]. However, Polish makes some additional fine-grained distinctions based on manner and speed of movement (e.g., jumping, fast vs. slow movement, motion with a change of direction). Whereas in Italian and English, verbs describing motion towards the speaker/listener form a distinct cluster. These preliminary analyses suggest that the collected semantic multi-arrangement data may support many other, fine-grained and in-depth lexical-typological analyses in future work, e.g., focusing on cross-lingual comparisons of the organisation of different semantic fields and examination of the most salient meaning dimensions underlying a given conceptual space.

4 Evaluation

Evaluation is focused on two types of representation architectures: static word embeddings (Bojanowski et al., 2017) and more recently proposed large pretrained encoders (Devlin et al., 2019). We compare their ability to capture word-level semantics across languages and domains of verb meaning. We also contrast the performance of language-specific BERT models with their massively multilingual counterpart (Devlin et al., 2019), and examine the impact of computing word-level representations *in context*, rather than by feeding items to a pretrained model *in isolation*.

Representation Models. We evaluate FASTTEXT (FT) as a representative non-contextualised word embedding model with proven representation capabilities on diverse NLP tasks (Mikolov et al., 2018) and coverage of 157 languages. For multi-word expressions, we compute their representations by averaging the vectors of their constituent words. We contrast the performance of FT vectors with the omnipresent state-of-the-art BERT model (Devlin et al., 2019). We derive word-level BERT representations of words and multi-word expressions in two different ways: (a) *in isolation* and (b) *in context*. In method (a), we follow the steps of Liu et al. (2019) by (1) feeding each item to the pretrained model in isolation, (2) averaging

[3]The easier, higher-IAA classes tend to include verbs whose meanings are more concrete (*boil, bake, grate*) or are organised along clear dimensions of meaning, e.g., based on increasing or decreasing intensity (negative and positive emotions, negative and positive rate of change); lower-IAA classes are usually bigger and hence more heterogeneous.

Models		Chinese		Japanese		Polish		Finnish		Italian	
	$k =$	*gold*	*optimal*	*gold*	*optimal*	*gold*	*optimal*	*gold*	*optimal*	*gold*	*optimal*
FASTTEXT		.314	.333	.250	.259	**.358**	**.377**	**.326**	**.386**	**.341**	**.389**
BERT											
(1) ISO		.246	.250	.221	.249	.190	.227	.249	.267	.205	.231
+WWM		-	-	.215	.244	-	-	-	-	-	-
+XXL		-	-	-	-	-	-	-	-	.205	.220
(2) CTX		**.333**	**.352**	.251	.253	.238	.265	.269	.306	.269	.270
+WWM		-	-	.237	.253	-	-	-	-	-	-
+XXL		-	-	-	-	-	-	-	-	.268	.300
M-BERT											
(1) ISO		.260	.284	.247	.264	.169	.216	.171	.211	.185	.196
(2) CTX		.264	.303	**.257**	**.271**	.216	.277	.200	.254	.227	.255

Table 5: Clustering results (F1 score) on Phase 1 classes, for the *optimal* clustering solution (highest F1 score) and with k clusters equal to the number of *gold* classes in each language (see Table 1). We report scores for (M-)BERT embeddings computed *in isolation* (ISO) and *in context* (CTX) (see §4).

the *H* hidden representations for each of the subword tokens constituting the item, and finally (3) taking the average of these subword representations to obtain the final d-dimensional representation ($d = 768$ in BERT-BASE). Again, following Liu et al. (2019), we average over all layers (12 with BERT-BASE). In (b), we encode word meaning in context of other words using external corpora[4] in the following way. First, we randomly sample N sentences containing each item in the corpus, then, we compute the item's representation in each of N sentential contexts (averaging over constituent subword representations and hidden layers as in steps (2)-(3) above), and finally average over the N sentential representations to obtain the final representation for each item.[5] We evaluate the uncased multilingual BERT model (M-BERT) (Devlin et al., 2019), pretrained on monolingual Wikipedia corpora of 102 languages, as well as language-specific pretrained BERT encoders released for ZH, JA (BERT-BASE with and without whole word masking (+WWM)), PL, FI, and IT (BERT-BASE and BERT-BASE-XXL trained on a larger Italian corpus), available in the Transformers repository (Wolf et al., 2019).[4]

4.1 Semantic Verb Clustering

First, we evaluate the models on semantic clustering, where the task is to group the starting verb sample (Table 1, **N verbs**) into clusters based on semantic similarity. For each vector collection, we apply the spectral clustering algorithm (Meila and Shi, 2001; Yu and Shi, 2003), shown to produce strong results in previous work on verb clustering (Sun et al., 2010; Scarton et al., 2014; Vulić et al., 2017), and evaluate the produced groupings against the Phase 1 classes in each language using standard clustering evaluation metrics, modified purity (MPUR) (i.e., mean precision of induced verb clusters) and weighted class accuracy (WACC), calculated as follows:

$$\text{MPUR} = \frac{\sum_{C \in \textbf{Clust}, n_{prev(C)}>1} n_{prev(C)}}{n_{test\_verbs}} \quad (1) \qquad \text{WACC} = \frac{\sum_{C \in \textbf{Gold}} n_{dom(C)}}{n_{test\_verbs}} \quad (2)$$

where (1) each cluster C from the set of all K_{Clust} automatically induced clusters **Clust** is associated with its prevalent Phase 1 class, and $n_{prev(C)}$ is the number of verbs in an induced cluster C appearing in that class (all other verbs are considered errors). n_{test_verbs} is the total number of test verbs, and singleton clusters ($n_{prev(C)} = 1$) are not counted. In Eq. (2), for each C from the set of Phase 1 classes **Gold** we identify the dominant cluster from the set of induced clusters which has most verbs in common with C ($n_{dom(C)}$). The metrics are combined into an F1 score, the balanced harmonic mean of MPUR and WACC.

Table 5 includes the results for the optimal number of clusters (highest F1), and for a fixed *k* equal to the number of gold truth classes. We observe several interesting patterns in the F1 scores. First, we note that FT vectors clearly outperform the BERT models in languages using the Latin script, IT, FI, PL,

[4]We provide details and URLs for the models and corpora used in this study in Appendices A and B.

[5]We tested different values of N (10, 100, 500) and due to negligible differences in scores only report results for $N = 100$.

Models	Chinese				Japanese				Polish				Finnish				Italian			
	THR	#1	#9	#2	THR	#1	#9	#2	THR	#1	#9	#2	THR	#1	#9	#2	THR	#1	#9	#2
FT	.261	.277	.111	**.425**	.067	.038	−.03	.022	**.254**	**.316**	.138	**.239**	**.248**	**.286**	**.307**	**.414**	**.326**	**.243**	.030	**.288**
BERT																				
(1) ISO	.187	.219	.041	.307	.054	.086	−.01	.230	.097	.205	−.02	.165	.112	.157	.110	.108	.097	.090	.008	.030
+WWM	-	-	-	-	.052	.164	−.06	.211	-	-	-	-	-	-	-	-	-	-	-	
+XXL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		.073	.083	.073	.016
(2) CTX	**.315**	**.344**	.231	.330	.067	.128	.064	.201	.124	.237	.130	.073	.188	.042	.245	.108	.134	.117	.085	.038
+WWM	-	-	-	-	.083	**.165**	.063	**.269**	-	-	-	-	-	-	-	-	-	-	-	-
+XXL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		.168	.128	**.123**	.079
M-BERT																				
(1) ISO	.161	.079	.166	.326	.098	.118	.060	.034	−.01	.201	.039	.024	.032	.028	.001	.137	.009	.043	−.22	.062
(2) CTX	.265	.305	**.236**	.213	**.101**	.116	**.098**	.128	.063	.093	**.166**	−.01	.076	.075	.174	−.07	.088	.146	.021	.004

Table 6: Word similarity evaluation results (Spearman's ρ) on the thresholded sets (THR) and three semantic domains, 'emotion' (#1), 'change' (#9), and 'cooking' (#2), in each language.

achieving the top three F1 scores overall (0.389, 0.386, 0.377).[6] In Chinese and Japanese, FT vectors surpass BERT embeddings *in isolation*, but are outperformed by BERT vectors computed *in context* (in ZH) and by the multilingual BERT in Japanese. The stronger performance of the massively multilingual model in Japanese and Chinese contrasts with the results in PL, FI and IT, where it mostly lags behind the language-specific counterparts. In terms of relative scores, we see that BERT and M-BERT embeddings computed over a number of sentential contexts consistently outperform their *in isolation* counterparts across all languages. On the other hand, we observe that whole word masking does not reliably improve clustering performance in Japanese, nor does using a larger training corpus in Italian (BERT-XXL).

Error Analysis. Manual inspection of the induced clusters reveals some common pitfalls and areas of difficulty. First, the evaluated models are largely oblivious to idiomatic meaning. In Polish and Italian, the FT model produces a cluster of 'possession' verbs (EN *have, give, lend, buy*), including the verbs *mieć* (PL,'to have'), *dać* and *dare* (PL/IT,'to give'). However, it also incorporates all phrasal verbs and multi-word expressions featuring these words, with meanings unrelated to the rest of the class: PL *mieć coś przeciw* ('to mind/object to'), *mieć nadzieję* ('to hope'), *mieć wpływ* ('to influence'), *dać klapsa* ('to spank'); IT *dare un'occhiata* ('to glance'). This is even more evident in Finnish, where a separate cluster of phrasal verbs with *olla* ('to be/have') emerges (e.g., *olla varuillaan*, 'beware', *olla peräisin*, 'originate', *olla samaa mieltä* 'agree'). Similarly, all Polish models produce a cluster of just reflexive verbs (e.g., *ślizgać się* ('to slide'), *cieszyć się* ('to rejoice'), *zdarzyć się* ('to happen')), regardless of discrepancies in meaning. In Italian, BERT models fall into the same trap, clustering reflexives regardless of their meaning (*informarsi* 'to inquire', *precipitarsi* 'to rush', *abbronzarsi* 'to tan'); however, FT vectors are more robust: we observe a separate cluster of movement verbs, with both reflexives and non-reflexives (*saltare* 'to jump', *precipitarsi* 'to rush', *andare* 'to go'), and of knowledge-related verbs (*informarsi* 'to inquire', *studiare* 'to study', *istruire* 'to instruct'). The attention to subword signal is apparent in clusters produced by BERT models. In languages using logographic scripts, this yields valid groupings, e.g., Japanese 再現する *saigen suru* 'to reproduce/reappear', 再生する *saisei suru* 'to reproduce', 再生利用する *saisei riyō suru* 'to recycle'. In Polish, however, *narzucać się* ('to impose') and *podrzucać* ('to toss'), and *polować* ('to hunt') and *malować* ('to paint') end up clustered together. While it could be argued that a weak semantic link (apart from the etymological one) exists between the first pair, the second pair has only coincidental orthographic overlap. Similarly, a semantically heterogeneous cluster of Italian verbs ending in *-lare* is produced (*coccolare* 'cuddle', *capitolare* 'capitulate', *scongelare* 'defrost').Whether computed *in context* or *in isolation*, BERT word-level representations capture a lot of subword- and surface-level information without fully capturing higher-level semantic signal, which negatively affects cluster quality.

4.2 Word Similarity

We compute Spearman's ρ correlation between the ranks of models' similarity scores and those of human judgments from Phase 2. To ensure reliability of the results, we perform evaluation on a thresholded

[6]Note that lower absolute scores are also due to the overlap in Phase 1 classes, while models perform hard clustering.

subset of each dataset focusing on classes with IAA above $\rho = 0.3$ (THR) (Table 1). We also compare the models' capacity to discriminate between related concepts within a narrow semantic domain and report scores on three semantic classes: 'emotion' (#1), 'change' (#9), 'cooking' (#2)[7] (Table 6).

The primacy of FT vectors in Polish, Finnish, and Italian is again conspicuous, while in Chinese and Japanese the pretrained encoders are in the lead, with noticeably lower FT performance recorded for Japanese than in the other languages. Results achieved on the THR sets repeat the patterns seen in the clustering task: contextualised variants of BERT embeddings (CTX) outperform those computed *in isolation* (ISO), and the language-specific encoders prove to capture richer semantics than the massively multilingual model - with the exception of Japanese, where contextualised M-BERT again achieves the top result (albeit noticeably lower than top THR scores in other languages). The relatively stronger M-BERT results on Japanese, as well as Chinese, illustrate the known unfavourable characteristic of multilingual pretraining with a subword vocabulary shared by 102 languages. The languages with scripts distinct from those of the majority of languages covered by M-BERT do not share their subwords with a large number of other languages, and their language-specific subwords constitute a large proportion of the total subword vocabulary; in consequence, the model can capitalise on this proliferation to produce higher-quality representations. Conversely, the representation quality is degraded for languages with very rich and productive morphology like Finnish or Polish, despite the availability of training data. This also applies to language-specific BERT models: given the same vocabulary capacity, morphologically rich languages have many words split into subwords and fewer full words represented in the vocabulary than analytic languages like Chinese or English.

To explore the potential of generating stronger word-level embeddings from BERT models, we investigated the impact of two parameters on lexical representation extraction: the number of hidden layers we average over (all 12 or first 8) and the inclusion of the special classification token ([CLS]) in the subword averaging step. Table 7 summarises the results for Polish and Finnish. We note that including the CLS token yields better lexical embeddings in both languages, however, whether averaging over 12 or 8 layers produces strongest results is language-specific. Notably, the top-performing Finnish BERT configuration ($\rho = 0.250$) outperforms FT embeddings ($\rho = 0.248$) on the THR set. While a full evaluation of all parameter configurations is beyond the scope of this paper, these findings suggest careful language-specific tuning of the extraction configuration is crucial to achieve optimal performance.

BERT		Polish				Finnish			
L	CLS	THR	#1	#9	#2	THR	#1	#9	#2
12	-	.097	.205	−.019	.165	.112	.157	.110	.108
	+	**.172**	**.228**	**.083**	**.202**	.227	.186	.151	.190
8	-	.086	.183	−.011	.141	.151	.223	.141	.160
	+	.142	.185	.081	.175	**.250**	**.234**	**.205**	**.280**

Table 7: Results for Polish and Finnish BERT *in isolation* vectors, averaged over all 12 or first 8 layers (*L*), with the **CLS** token (**+**) or without it (**-**).

Although more variation in terms of primacy of one model variant over the other is expected on the semantic classes due to smaller dataset size, the general pattern whereby computing BERT embeddings by averaging *N* contextualised representations boosts performance applies in 72% of cases. Additional observations can be drawn from results on individual domains. Class #9, including verbs describing change in size or speed (e.g., *accelerate, increase, shrink*), is especially rich in synonyms and antonyms. Due to antonyms' high semantic overlap they are often confused with synonyms by distributional models learning purely from patterns of occurrences in raw text. This effect also emerges in our results, where performance on this class is the lowest for most model configurations and languages. Finnish is the exception, possibly due to the slightly broader coverage of this class, which also incorporates verbs of being and existence (Table 3), with smaller proportion of antonymous pairs. Interestingly, this class is where the multilingual model outperforms the language-specific counterparts in ZH, JA, PL. In Italian, where class #9 has only 23 members, most of which stand in antonymous relations (e.g., *crescere - decrescere* 'increase - decrease', *aumentare - diminuire* 'rise - drop', *iniziare - finire* 'begin - finish'), the BERT model trained on the larger corpus is the most robust. Results on this class illustrate that semantic areas which are easier for humans to reason about are not necessarily less challenging for models.

An area where greater ease of human judgment is reflected in relatively higher model performance is

[7]We carried out evaluation on all classes but report selected results (on highest IAA classes cross-lingually) for brevity.

the domain of cooking verbs in ZH, FI, JA, where the highest overall scores are recorded (>0.4 for FT in ZH and FI), and the top model scores in Japanese (BERT). While we do not report all class-specific correlations for brevity, they reveal further interesting patterns as to the semantic domains which prove easiest for models to accurately capture. In IT, we record highest model correlations for verbs of communication and destruction (top scores >0.4 (FT)), while verbs of physical violence are the domain with highest correlations in PL (>0.3 FT), FI (>0.4 FT) and ZH (>0.4 BERT). In Japanese, the best result overall is achieved on verbs of cognition (0.276 BERT), followed by verbs of physiological processes with >0.2 correlations scored by the contextualised BERT models. Similar analyses on specific semantic domains can help identify strengths and deficiencies of different types of embeddings and highlight the areas of meaning which pose challenges across languages, guiding further developments in representation learning.

4.3 Main Observations

Our evaluation revealed the dataset to be a challenging benchmark, and provided a number of insights into the potential of the evaluated models to capture verbal lexical semantics across languages.

- Overall, model performance across tasks shows a split pattern: the pretrained encoders surpass static word embeddings in Chinese and Japanese, but are outperformed by FASTTEXT by a significant margin in languages using the Latin script, Polish, Finnish, and Italian. There is potential to derive higher-quality word-level BERT embeddings in those languages through careful selection of language-specific lexical representation extraction parameters.

- BERT word-level embeddings derived by averaging over *N* occurrences in context prove predominantly stronger than those obtained by feeding words into the pretrained model in isolation, with more variation observed in the case of the smaller semantic sets.

- The results achieved on the thresholded datasets show a clear advantage of monolingual pretraining over the massively multilingual pretraining - with the exception of Japanese, where M-BERT achieves the top results, as is the case in semantic clustering.

- Error analysis revealed that clustering performance of the pretrained encoders suffers due to the primacy given to low-level subword signal over the high-level semantic information, while an important area of difficulty for all models in the lexical similarity task is the problem of teasing apart antonymous and synonymous word pairs.

5 Conclusion and Future Work

We presented the first large-scale multilingual evaluation resource, constructed via spatial arrangement and targeting verb semantics in Chinese, Japanese, Polish, Finnish, and Italian. It includes semantic classes and fine-grained pairwise lexical similarity scores, which we release with this paper. The dual nature and vast coverage of the dataset enables evaluation of representation models on two tasks, semantic clustering and word similarity, and focused probing analyses on specific semantic domains, revealing aspects of verbal meaning which elude models' representation capacity. The low overall model performance indicates that estimating similarity between a large number of semantically proximate concepts linked by fine-grained relations is a challenging task. In future work, we will use the spatial arrangement data for in-depth analyses of cross-lingual typological variation and model evaluation on fine-grained semantic clusters from Phase 2 to explore the potential for (semi-)automatic creation of verb classes and semantic resources in languages where those are still lacking. We will also evaluate cross-lingual representation learning algorithms on mapped cross-lingual verb similarity datasets for all language pairs created in this project.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions. We would like to thank Daisuke Kawahara and Qianchu Liu for their kind help and contributions to the creation of the dataset. This work is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909) and the Economic and Social Research Council [PhD Award Number ES/J500033/1] (OM).

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL-HLT*, pages 19–27.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of EMNLP*, pages 278–289.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Daniel Casasanto. 2008. Similarity and proximity: When does close in space mean close in mind? *Memory & Cognition*, 36(6):1047–1056.

Radoslaw M. Cichy, Nikolaus Kriegeskorte, Kamila M. Jozwik, Jasper J.F. van den Bosch, and Ian Charest. 2019. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gökhan Ercan and Olcay Taner Yıldız. 2018. Anlamver: Semantic model evaluation dataset for turkish-word similarity and relatedness. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3819–3836.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database and Some of Its Applications*. MIT Press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Peter Gärdenfors. 2004. *Conceptual Spaces: The Geometry of Thought*. MIT press.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of ICML*.

Junjie Huang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, and Maosong Sun. 2019. COS960: A Chinese word similarity dataset of 960 word pairs. *CoRR*, abs/1906.00247.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Proceedings of SEMEVAL*, pages 290–299.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC*, pages 1027–1032.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings EMNLP-IJCNLP*, pages 2779–2795.

Nikolaus Kriegeskorte and Marieke Mur. 2012. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3:245.

George Lakoff and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, volume 4. University of Chicago Press.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *CoRR*, abs/1508.00106.

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of CoNLL*, pages 33–43.

Olga Majewska, Diana McCarthy, Jasper van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, and Anna Korhonen. 2020. Spatial multi-arrangement for clustering and multi-way similarity dataset construction. In *Proceedings of LREC*, pages 5751–5760.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220.

Marina Meila and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *Proceedings of AISTATS*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*, pages 3111–3119.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of LREC*, pages 52–55.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of ACL*, pages 8093–8104.

Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A. Bandettini, and Nikolaus Kriegeskorte. 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, 4:128.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Piotr Rychlik. 2018. SimLex-999 for Polish. In *Proceedings of LREC*.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making sense of language-specific BERT models. *CoRR*, abs/2003.02912.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *SCIENCE CHINA Technological Sciences*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Yuya Sakaizawa and Mamoru Komachi. 2018. Construction of a Japanese word similarity dataset. In *Proceedings of LREC*, pages 948–951.

Carolina Scarton, Lin Sun, Karin Kipper-Schuler, Magali Sanches Duran, Martha Palmer, and Anna Korhonen. 2014. Verb clustering for Brazilian Portuguese. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 25–39.

Lin Sun, Anna Korhonen, Thierry Poibeau, and Cédric Messiant. 2010. Investigating the cross-linguistic potential of VerbNet style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1056–1064.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *CoRR*, abs/1912.07076.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of EMNLP*, pages 2546–2558.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *arXiv preprint arXiv:2003.04866*.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of EMNLP-IJCNLP*, pages 5721–5727.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.

Dongqiang Yang and David M. W. Powers. 2006. Verb similarity on the taxonomy of WordNet. In *Proceedings of the 3rd International WordNet Conference (GWC-06)*, pages 121–128.

Stella X. Yu and Jianbo Shi. 2003. Multiclass spectral clustering. In *Proceedings of ICCV*, page 313.

A Representation Models

We provide URLs to the models used in this study in Table 8 below. For all languages, we used the pre-trained uncased BERT-base models. We also evaluate 300-dimensional `fastText` vectors (Mikolov et al., 2018), trained on Common Crawl and Wikipedia data of each language using an extension of the CBOW `word2vec` model (Mikolov et al., 2013) with position-weights over 10 training epochs, with character n-grams of length 5, window size of 5, and 10 negative examples.

Language	Model	URL
Chinese	BERT	https://huggingface.co/bert-base-chinese
Finnish	BERT	https://huggingface.co/TurkuNLP/bert-base-finnish-uncased-v1
Italian	BERT	https://huggingface.co/dbmdz/bert-base-italian-uncased
	+XXL	https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased
Japanese	BERT	https://huggingface.co/cl-tohoku/bert-base-japanese
	+WWM	https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
Polish	BERT	https://huggingface.co/dkleczek/bert-base-polish-uncased-v1
Multilingual	BERT	https://huggingface.co/bert-base-multilingual-uncased
all	FT	https://fasttext.cc/docs/en/crawl-vectors.html

Table 8: Links to the models used in this study. For each language, we used the uncased BERT-base model(s) (including the variant with whole word masking (+WWM) for Japanese and the XXL Italian BERT-base model trained on a larger (81GB) corpus) and 300-dimensional `fastText` (FT) vectors available for that language.

B External Corpora

The corpora used to extract sentential contexts for the *in context* BERT word embeddings are listed below (Table 9). We randomly sampled 1 million sentences of maximum sequence length 512 from each monolingual corpus.

Language	Corpus	URL	Word segmenter
Chinese	United Nations Parallel Corpus	http://opus.nlpl.eu/UNPC.php	https://github.com/fxsjy/jieba
Finnish	Europarl	http://opus.nlpl.eu/Europarl.php	-
Italian	Europarl	http://opus.nlpl.eu/Europarl.php	-
Japanese	Polyglot Wikipedia	https://sites.google.com/site/rmyeid/projects/polyglot?authuser=0	https://github.com/Kensuke-Mitsuzawa/JapaneseTokenizers
Polish	Europarl	http://opus.nlpl.eu/Europarl.php	-

Table 9: Links to the external corpora used for extraction of N sentences for computing BERT representations *in context* and the word segmenters used, where appropriate.

C WALS Features

Table 10 lists the morphological, syntactic and lexical typological features from the World Atlas of Language Structures (WALS) (`https://wals.info`) used in cross-lingual comparisons in Section 3.2 (Table 2), selected based on the availability of corresponding entries for the languages in our sample.

ID	Feature	ID	Feature
26A	Prefixing vs. Suffixing in Inflectional Morphology	86A	Order of Genitive and Noun
29A	Syncretism in Verbal Person/Number Marking	87A	Order of Adjective and Noun
33A	Coding of Nominal Plurality	88A	Order of Demonstrative and Noun
36A	The Associative Plural	89A	Order of Numeral and Noun
40A	Inclusive/Exclusive Distinction in Verbal Inflection	90A	Order of Relative Clause and Noun
44A	Gender Distinctions in Independent Personal Pronouns	91A	Order of Degree Word and Adjective
45A	Politeness Distinctions in Pronouns	92A	Position of Polar Question Particles
46A	Indefinite Pronouns	95A	Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase
47A	Intensifiers and Reflexive Pronouns	96A	Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun
48A	Person Marking on Adpositions	97A	Relationship between the Order of Object and Verb and the Order of Adjective and Noun
49A	Number of Cases	100A	Alignment of Verbal Person Marking
50A	Asymmetrical Case-Marking	101A	Expression of Pronominal Subjects
51A	Position of Case Affixes	102A	Verbal Person Marking
52A	Comitatives and Instrumentals	103A	Third Person Zero of Verbal Person Marking
53A	Ordinal Numerals	104A	Order of Person Markers on the Verb
64A	Nominal and Verbal Conjunction	105A	Ditransitive Constructions: The Verb 'Give'
69A	Position of Tense-Aspect Affixes	107A	Passive Constructions
70A	The Morphological Imperative	112A	Negative Morphemes
71A	The Prohibitive	115A	Negative Indefinite Pronouns and Predicate Negation
72A	Imperative-Hortative Systems	116A	Polar Questions
74A	Situational Possibility	129A	Hand and Arm
75A	Epistemic Possibility	130A	Finger and Hand
76A	Overlap between Situational and Epistemic Modal Marking	138A	Tea
80A	Verbal Number and Suppletion	143A	Order of Negative Morpheme and Verb
81A	Order of Subject, Object and Verb	143E	Preverbal Negative Morphemes
82A	Order of Subject and Verb	143F	Postverbal Negative Morphemes
83A	Order of Object and Verb	143G	Minor morphological means of signaling negation
84A	Order of Object, Oblique, and Verb	144A	Position of Negative Word With Respect to Subject, Object, and Verb
85A	Order of Adposition and Noun Phrase		

Table 10: WALS typological features considered in cross-lingual comparisons.